Switching the Loss Reduces the Cost in Batch Reinforcement Learning

Alex Ayoub [1] **Kaiwen Wang** [2] **Vincent Liu** [1] **Samuel Robertson** [1] **James McInerney** [3] **Dawen Liang** [3] **Nathan Kallus** [3 2] **Csaba Szepesvári** [1]

Abstract

We propose training fitted Q-iteration with log-loss (FQI-LOG) for batch reinforcement learning (RL). We show that the number of samples needed to learn a near-optimal policy with FQI-LOG scales with the accumulated cost of the optimal policy, which is zero in problems where acting optimally achieves the goal and incurs no cost. In doing so, we provide a general framework for proving *small-cost* bounds, i.e. bounds that scale with the optimal achievable cost, in batch RL. Moreover, we empirically verify that FQI-LOG uses fewer samples than FQI trained with squared loss on problems where the optimal policy reliably achieves the goal.

1. Introduction

In batch reinforcement learning (RL), also known as offline RL, the goal is to learn a good policy from a fixed dataset. A standard approach in this setting is fitted Q-iteration (FQI) (Ernst et al., 2005), which iteratively obtains a sequence of value functions by fitting the next value function to a target that is based on the data and the previously obtained value function. In this work we propose a simple and principled improvement to FQI, using *log-loss* (FQI-LOG), which is applicable when the returns along trajectories lie in a bounded interval. We prove that the number of samples the new method requires to learn a near-optimal policy scales with the cost of the optimal policy, leading to a so-called *small-cost* bound, the RL analogue of a small-loss bound in supervised learning. Such bounds predict improved sample efficiency in goal oriented RL tasks where the goal is reliably achievable and the cost is set up to penalize failure in achieving the goal; a prediction we validate empirically. We highlight that FQI-LOG is the first *computationally efficient* batch RL algorithm to achieve a small-cost bound, provided that a regression oracle is available; a condition that can be met, e.g., when logit models are used in FQI-LOG.

Most earlier works in batch RL focused on algorithms that achieve the optimal worse-case dependence on the number of samples required to learn a near-optimal policy (Munos, 2003; Antos et al., 2007; Chen & Jiang, 2019; Xie & Jiang, 2021). The only work prior to ours that is known to *adaptively* improve sample efficiency when facing a task with near-zero optimal cost is due to Wang et al. (2023), who obtain small-cost bounds for finite-horizon batch RL problems but using the so-called "distributional RL" approach. In this approach, one solves the regression problems arising when estimating the distribution of a policy's accumlated cost. The inspiration for our work comes from this work, combined with the insights of Abeille et al. (2021); Foster & Krishnamurthy (2021) who showed that, in the simpler bandit problems, log-loss alone is sufficient for obtaining small-cost bounds.

Why log-loss gives a small-cost advantage is subtle. When used with an unrestricted model class (think: finite state-action space, "tabular learning"), both log- and squared losses achieve small cost bounds because they share the same minimizers, which predict the empirical mean of responses for all inputs. However, when the model class is restricted (the only practical case for large problems), log-loss and squared loss will trade off errors at the various inputs differently. Specifically, with log-loss, the penalty for predicting a value far from an observed mean increases rapidly as the observed mean gets close to the boundary of its range, an effect that is absent with squared loss. Consequently, log-loss will favor predictors that fit well to those observed values that are near the boundary of the range, making the learning process disregard large variance observed values, stabilizing learning, a property not shared when a squared loss is used. As a result, as we show, under suitable additional technical assumptions, log-loss based FQI achieves small-cost bounds, a property that is not shared when squared loss is used with FQI.

The main contributions of this work can be summarized as: *(i)* We propose training FQI with log-loss and prove it enjoys a small-cost bound. This is the first efficient batch RL algorithm that achieves a small-cost bound. When showing

[1]University of Alberta [2]Cornell University [3]Netflix, Inc.. Correspondence to: Alex Ayoub <aayoub14k@outlook.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

this result, we make two technical contributions that may be of independent interest: *(ii)* We show that the Bellman optimality operator is a contraction with respect to the Hellinger distance, a result; *(iii)* We present a general result that decomposes the suboptimality gap of the value of a greedy policy induced by some value function, f, into the product of a small-cost term and the pointwise triangular deviation of f from $q^\star$, the value function of the optimal policy.

2. Preliminaries

In this section we review some definitions and concepts of Markov Decision Processes (MDPs) and define the notation used. Readers unfamiliar with the basic theory of MDPs are recommended to consult the book of Bertsekas (2019), or that of Szepesvári (2010). All results mentioned in this section can be found in these works.

An infinite-horizon discounted Markov Decision Process (MDP) is given by a tuple $M = (\mathcal{S}, \mathcal{A}, P, c, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, P is a transition function, $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a cost function and $\gamma \in [0, 1)$ is a discount factor. We only consider MDPs with finite action spaces. Furthermore, for simplicity, we assume that the state space is finite. Among other things,[1] this allows writing the transition function as $P : \mathcal{S} \times \mathcal{A} \to \mathcal{M}_1(\mathcal{S})$, where $\mathcal{M}_1(\mathcal{S})$ denotes the set of probability distributions over $\mathcal{S}$. Since the set $\mathcal{S}$ is finite, any element of $\mathcal{M}_1(\mathcal{S})$ can and will be identified with its probability mass function. The notation $\mathcal{M}_1(\mathcal{X})$ will be used in the same way to denote the set of probability distributions over an arbitrary finite set $\mathcal{X}$ and we will perform the same identification.

A (general) policy $\pi = (\pi_h)_{h=1}^\infty$ is a sequence of functions $\pi_h : (\mathcal{S} \times \mathcal{A})^{h-1} \times \mathcal{S} \to \mathcal{M}_1(\mathcal{A})$. Fixing the start state s, a policy π induces a distribution $\mathbb{P}_{\pi,s}$ over trajectories $S_1, A_1, C_1, S_2, A_2, C_2, \ldots$, where $S_1 = s$, $A_1 \sim \pi_1(S_1)$, $C_1 = c(S_1, A_1)$, $S_2 \sim P(S_1, A_1)$, $A_2 \sim \pi_2(S_1, A_1, S_2)$: the policy is used to govern the selection of actions, while the transition dynamics of the MDP governs the evolution of the states in response to the chosen actions. We will also need *stationary Markov* policies, where the choice of the action in any timestep h only depends on the last state visited. Thus, such policies can be identified with a map $\pi : \mathcal{S} \to \mathcal{M}_1(\mathcal{A})$, an identification which we will employ in what follows.

The expected total discounted cost over trajectories starting in s quantifies the policy's performance when initialized in state s. We collect these expectations in the *state-value function* of π, $v^\pi : \mathcal{S} \to \mathbb{R}$, which is defined by $v^\pi(s) = \mathbb{E}_{\pi,s}[\sum_{h=1}^\infty \gamma^{h-1} C_h]$, where $\mathbb{E}_{\pi,s}$ is the expectation operator corresponding to $\mathbb{P}_{\pi,s}$. *For convenience, throughout this paper we assume that* costs are normalized *so that the sum of discounted costs along any trajectory satisfies*

$$0 \le \sum_{h=1}^\infty \gamma^{h-1} C_h \le 1\,. \tag{1}$$

By appropriately rescaling the costs, this constraint can always be satisfied (when the state space is not finite, one needs that the above infinite sums are uniformly bounded to be able to do this).

The *action-value function* of π, $q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, is defined as

$$q^\pi(s, a) = c(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, a) v^\pi(s')\,,$$

where $(s, a) \in \mathcal{S} \times \mathcal{A}$ and, by abusing notation, we use $P(s'|s, a)$ to denote the probability of landing in state s' when action a is taken in state s. For a stationary Markov policy π, the state- and action-value functions are related by the identity $v^\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) q^\pi(s, a)$. Here, and in what follows, abusing notation once again, $\pi(a|s)$ denotes the probability that is assigned by $\pi(s)$ to action $a \in \mathcal{A}$.

The *optimal policy* is defined as any policy $\pi^\star$ that satisfies $v^{\pi^\star}(s) = \min_\pi v^\pi(s)$ simultaneously for all $s \in \mathcal{S}$. Such a policy exists in our case. We define the *optimal state-value function* as $v^\star = v^{\pi^\star}$ and the *optimal action-value function* as $q^\star = q^{\pi^\star}$. Any policy that is *greedy with respect to* $q^\star$, i.e., at state s selects only actions a that minimize $q^\star(s, \cdot)$, is guaranteed to be optimal. Furthermore, the optimal action-value function $q^\star$ satisfies the *Bellman optimality equation* $q^\star = \mathcal{T} q^\star$, where $\mathcal{T} : \mathbb{R}^{\mathcal{S} \times \mathcal{A}} \to \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$ is the *Bellman optimality operator*, defined via

$$(\mathcal{T} f)(s, a) = c(s, a) + \gamma \sum_{s' \in S} P(s'|s, a) \min_{a' \in \mathcal{A}} f(s', a'), \tag{2}$$

for $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$.

We will find it helpful to use a shorthand for the function $s \mapsto \min_{a \in \mathcal{A}} f(s, a)$ appearing above. For $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, define $f^\wedge : \mathcal{S} \to \mathbb{R}$ by

$$f^\wedge(s) = \min_{a \in \mathcal{A}} f(s, a)\,, \qquad s \in \mathcal{S}\,.$$

With this notation, we have that for $(s, a) \in \mathcal{S} \times \mathcal{A}$

$$(\mathcal{T} f)(s, a) = c(s, a) + \gamma \sum_{s' \in S} P(s'|s, a) f^\wedge(s')\,.$$

Finally, we use π_f to denote a greedy policy induced by f: $\pi_f(s) = \arg\min_{a \in \mathcal{A}} f(s, a)$. When there are multiple such policies, we choose one in an arbitrary (systematic) manner to make π_f well-defined.

[1] We assume that the state space $\mathcal{S}$ is finite solely for exposition. This allows us to simplify the presentation of our analysis and focus on the most salient details of the proof, avoiding the cumbersome measure-theoretic notation required to reason about infinite sets.

3. Problem Definition: Batch RL

In this work, we consider batch reinforcement learning problems, where a learner is given a sequence of data points $D_n = \{(S_i, A_i, C_i, S_i')\}_{i=1}^n$ such that $S_i, S_i' \in \mathcal{S}$, $A_i \in \mathcal{A}$. and $C_i \in \mathbb{R}$. Importantly, the learner has no access to P or c. The learner's goal is to find a policy π such that executing the policy π from an initial state S_1 drawn randomly from some distribution $\eta_1 \in \mathcal{M}_1(\mathcal{S})$ results in an expected total discounted cost exceeding the smallest possible such value with as little as possible. Formally, the learner is evaluated by comparing the value $\bar{v}^\pi = \langle \eta_1, v^\pi \rangle = \sum_{s\in\mathcal{S}} \eta_1(s) v^\pi(s)$, where $\langle \cdot, \cdot \rangle$ denotes the standard inner product, of the policy π that it returns with the smallest possible such value $\bar{v}^\star$, which, thanks to our earlier definitions is easily seen to satisfy $\bar{v}^\star = \langle \eta_1, v^\star \rangle$. Our algorithm does not need to know η_1.

The data is assumed to satisfy the following assumption:

Assumption 3.1 (Data/MDP properties)**.** We have that $D_n = \{(S_i, A_i, C_i, S_i')\}_{i=1}^n$ is a sequence of independent, identically distributed random variables such that $(S_i, A_i) \sim \mu$, $C_i = c(S_i, A_i)$ and $S_i' \sim P(S_i, A_i)$. Furthermore, Equation (1) holds and c is a nonnegative function.

In this assumption the constraints on C_i and S_i' will help the learner to discover some information about the costs and the transition structure of the MDP. The independence assumption is made to simplify the analysis. The distribution μ will be further restricted to be "sufficiently exploratory". To state this assumption, the following definition will be useful:

Definition 3.1 (Admissible distribution)**.** We say a distribution $\nu \in \mathcal{M}_1(\mathcal{S} \times \mathcal{A})$ is admissible in MDP M if there exists $h \geq 1$ and a nonstationary policy π such that $\nu(s, a) = \mathbb{P}_{\pi,\eta_1}(S_h = s, A_h = a)$.

Note that what constitutes an admissible distribution depends on η_1. With this, the assumption that constrains μ is as follows:

Assumption 3.2 (Finite concentrability coefficient)**.** There exists $C < \infty$ such that for all admissible distributions ν of M, it holds that

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{\nu(s,a)}{\mu(s,a)} \leq C\,. \tag{3}$$

Note that if μ is positive over $\mathcal{S} \times \mathcal{A}$, the assumption is satisfied. By only considering admissible distributions, we allow μ to be "concentrated" on states that are "relevant" in the sense that they are visited by some policy in some time step with a large probability when starting from η_1.

In addition to having access to the data, we will also assume that the learner is given access to a set of functions, $\mathcal{F}$ that map state-action pairs to reals: $\mathcal{F} \subset \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$. Ideally, the set $\mathcal{F}$ allows the learner to reason about the optimal action-value function. To make this possible, the following assumptions are made on $\mathcal{F}$:

Assumption 3.3 (Realizability)**.** $q^\star \in \mathcal{F}$.

Assumption 3.4 (Completeness)**.** For all $f \in \mathcal{F}$ we have that $\mathcal{T}f \in \mathcal{F}$.

Assumptions 3.1 to 3.4 are commonly made, in various forms, when analyzing fitted Q-iteration (Farahmand, 2011; Pires & Szepesvári, 2012; Chen & Jiang, 2019). Assumption 3.2 ensures that all admissible distributions are covered by the exploratory distribution μ, i.e., that "μ is sufficiently exploratory". Assumption 3.3 ("realizability") guarantees that the optimal action-value function, our ultimate target, lies in our function class. Assumption 3.4 states that the function class $\mathcal{F}$ is closed under the Bellman optimality operator $\mathcal{T}$. When $\mathcal{F}$ is closed, completeness is easily seen to imply realizibility (see also footnote 10 of Chen & Jiang (2019)). Note that Assumption 3.4 is necessary, this is due to a result by Foster et al. (2021) which states that assuming both a finite concentrability coefficient and a realizable function class are not sufficient for sample efficient batch value function approximation. For a more detailed discussion of the last three assumptions, we refer the reader to Sections 4 and 5 of Chen & Jiang (2019).

Research question As is well known, under the above assumptions, and assuming that a regression oracle is available to find the empirical minimizer of regression problems defined over $\mathcal{F}$ with the squared loss, the so-called fitted Q-iteration (FQI) algorithm (Ernst et al., 2005; Riedmiller, 2005; Antos et al., 2007) produces a policy such that with high probability $\bar{v}^\pi \leq \bar{v}^\star + \tilde{O}(\sqrt{CN/n})$, where N is a measure that characterizes the "richness" of $\mathcal{F}$ and $\tilde{O}$ hides logarithmic factors (Antos et al., 2007). The main question investigated in this paper is whether this result can be improved to $\bar{v}^\pi \leq \bar{v}^\star + \tilde{O}(\sqrt{CN\bar{v}^\star/n}) + O(1/n)$. The significance of such *small cost* results is that the same data can produce a significantly better policy when $\bar{v}^\star$ is near zero (note that $\bar{v}^\star \geq 0$). Alternatively, the number of samples required to achieve a given level of suboptimality can be significantly smaller if an algorithm satisfies a small-cost bound.

Additional notation For $n \in \mathbb{N}$, let $[n]$ denote the set $\{1, 2, \ldots, n\}$. Let $\mathbb{P}_{\pi,\eta_1}$ denote the distribution induced over random trajectories by following policy π after an initial state is sampled from η_1. For $h \in \mathbb{N}$, we let $\eta_h^\pi(s)$ be the probability that state s is observed at timestep h under $\mathbb{P}_{\pi,\eta_1}$, such that $\eta_h^\pi(s) = \mathbb{P}_{\pi,\eta_1}(S_h = s)$. We also define $\eta_h^\star(s) = \eta_h^{\pi^\star}(s)$. For $g : \mathcal{X} \to \mathbb{R}$, $\nu \in \mathcal{M}_1(\mathcal{X})$, and $p \geq 1$, we define the semi-norm $\|\cdot\|_{p,\nu}$ via $\|g\|_{p,\nu}^p = \int |g|^p d\nu$. We adopt standard big-oh notation and write $f = \tilde{\mathcal{O}}(g)$

to denotes that f dominates g up to polylog factors, i.e., $f = \mathcal{O}(g \max\{1, \text{polylog}(g)\})$. Finally, we use $f \wedge g$ to denote $\min\{f, g\}$.

4. FQI-LOG: Fitted Q-Iteration with log-loss

The proposed algorithm, FQI-LOG, which is described in Algorithm 1, is based on the earlier mentioned fitted Q-iteration algorithm (FQI) (Ernst et al., 2005; Riedmiller, 2005; Antos et al., 2007). Given a batch dataset, FQI iteratively produces a sequence of k approximations $f_1, \ldots, f_k$ to the action-value function $q^\star$. At iteration $j \in [k]$, the algorithm computes f_j by minimizing the empirical loss using targets computed with the help of f_{j-1}, the estimate produced in the previous iteration. The targets are constructed such that the regression function for a fixed estimate $f \in \mathcal{F}$ is $\mathcal{T} f$. The main difference between the proposed method and the most common variant of FQI is our use of log-loss,

$$\ell_{\log}(y, \tilde{y}) = \tilde{y} \log \frac{1}{y} + (1 - \tilde{y}) \log \frac{1}{1 - y}\,, \qquad (4)$$

to measure the deviation between a prediction ($y \in [0, 1]$) and a target ($\tilde{y} \in [0, 1]$), where to allow $y, \tilde{y} \in \{0, 1\}$, we use $0 \cdot \log 0 = 0$. The restriction that $\tilde{y} \in [0, 1]$ means that $\ell_{\log}(\cdot, \tilde{y})$ is convex. In our algorithm the first argument of $\ell_{\log}$ is a predicted value $f(S_i, A_i)$ with $f \in \mathcal{F}$. Since this needs to also belong to $[0, 1]$, running FQI-LOG requires the range of all functions in $\mathcal{F}$ to lie in $[0, 1]$.

Note that previous work on FQI employed the squared loss $\ell_{\text{sq}} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, defined as $\ell_{\text{sq}}(y, \tilde{y}) = (y - \tilde{y})^2$, instead of the log-loss $\ell_{\log}$. Both squared loss and log-loss have the property that for a random variable X taking values in $[0, 1]$ with probability one, $\mathbb{E}[X] = \arg\min_{m \in \mathbb{R}} \mathbb{E}[\ell(m, X)]$, where ℓ is either $\ell_{\log}$ or ℓ_{sq}. In particular, if $\mathcal{F} = [0, 1]^{\mathcal{S} \times \mathcal{A}}$, both log-loss and squared loss will give rise to the same sequence $f_1, \ldots, f_k$.

The differences between log-loss and squared loss are made apparent when $\mathcal{F} \neq [0, 1]^{\mathcal{S} \times \mathcal{A}}$. In this case, when $\tilde{y}$ is near 0 or 1, $\ell_{\log}(\cdot, \tilde{y})$ increases rapidly as y deviates from $\tilde{y}$. As such, when errors need to be traded off at different inputs, log-loss will end up favoring predictors that predict values closer to observed targets when the targets are near 0 (or 1), and will put less weight on observed targets in the middle of the $[0, 1]$ range. When a true value lies near 0 (or 1), the observed value (bound to the range $[0, 1]$) must be closer to the true value, which means that the observed value is also close to 0 (or 1). While it may happen that an observed value is close to 0 (or 1) while its mean is far from it, this is rare: For this to happen, the observed value has to have a large variance. As such, favoring to predict observed values near the 0 or 1 as opposed to paying equal attention to all datapoints (which is what the squared loss based predictors do) is beneficial, and, in particular it pays off when some of the true targets are near 0 (or 1): The situation that arises when the optimal cost is near zero.

Algorithm 1 FQI-LOG

Input: A dataset $D_n = \{(S_i, A_i, C_i, S_i')\}_{i \in [n]}$, a function class $\mathcal{F} \subseteq [0, 1]^{\mathcal{S} \times \mathcal{A}}$ and $k \in \mathbb{N}$.
Pick f_0 arbitrarily from $\mathcal{F}$
for $j = 1$ **to** k **do**
  $f_j \leftarrow \arg\min_{f \in \mathcal{F}} \sum_{i=1}^n \ell_{\log}\left(f(S_i, A_i), C_i + \gamma f_{j-1}^{\wedge}(S_i')\right)$
end for
Return: π_{f_k}.

The motivation to switch to log-loss is due to Foster & Krishnamurthy (2021) who studied the problem of learning a near-optimal policy in contextual bandits, both in the batch and the online settings. They noticed that switching to log-loss from squared loss allows bounding the suboptimality of the policy found, say in the batch setting after seeing n contexts, via a term that scales with $\sqrt{\bar{v}^\star / n} + 1/n$. This is an improvement from the usual $\sqrt{1/n}$ bound derived when analyzing squared loss, which is worst-case in nature. For log-loss, a significant speedup to $1/n$-type convergence is achieved when $\bar{v}^\star$, the expected cost of using the optimal policy, is small (cf. Section 3.1 of their paper). They complemented the theory with convincing empirical demonstrations. Our results take a similar form. While we reuse some of their results and techniques, our analysis deviates significantly from theirs. In particular, our analysis must be adapted to handle the multistage structure present in RL and to avoid an unnecessary dependence on the actions.

The astute reader may wonder whether switching to log-loss is really necessary for achieving small-cost bounds. As it turns out, the switch is necessary, as attested to by an example constructed by Foster & Krishnamurthy (2021). In this example, in contrast to log-loss, squared loss is shown to be unable to take advantage of small optimal costs (cf. Theorem 2 of Foster & Krishnamurthy (2021)).

5. Theoretical Results

In this section, we present our main theoretical contribution, the first *small-cost* bound for an efficient algorithm in batch RL.

Theorem 5.1. *Given a dataset $D_n = \{(S_i, A_i, C_i, S_i')\}_{i=1}^n$ with $n \in \mathbb{N}$ and a finite function class $\mathcal{F} \subseteq [0, 1]^{\mathcal{S} \times \mathcal{A}}$ that satisfy Assumptions 3.1 to 3.4, it holds with probability $1 - \delta$ that the suboptimality gap $g = \bar{v}^{\pi_k} - \bar{v}^\star$ of the output policy of FQI-LOG after k iterations, $\pi_k = \pi_{f_k}$, satisfies*

$$g \leq \tilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^2}\left(\sqrt{\frac{\bar{v}^\star C N}{n}} + \frac{C N}{(1-\gamma)^2 n} + \gamma^k\right)\right),$$

where $N = \log(|\mathcal{F}|/\delta)$ and C is defined in Assumption 3.2.

The full statement of Theorem 5.1, including lower order terms, can be found in Appendix B along with its proof. Compared to prior error bounds for FQI (Antos et al., 2007; 2008; Munos & Szepesvári, 2008; Farahmand, 2011; Lazaric et al., 2012; Chen & Jiang, 2019), to the best of our knowledge, Theorem 5.1 is the first that contains the instance-dependent optimal cost $\bar{v}^\star$. This makes Theorem 5.1 a small-cost bound, also referred to as a *first-order* (Freund & Schapire, 1997; Neu, 2015) or *small-loss* (Lykouris et al., 2022; Wang et al., 2023) bound in the learning theory literature. All previous results for FQI obtain an error bound independent of $\bar{v}^\star$, and cannot be made to scale with $\bar{v}^\star$ due to their use of squared loss (see Theorem 2 of Foster & Krishnamurthy (2021), mentioned earlier). Finally, we highlight that Theorem 5.1 is the first small-cost bound for a batch RL algorithm that is *computationally efficient* when efficient regression oracles are available, as is the case when $\mathcal{F}$ is the set of logit models with weights bounded in 2-norm. While technically this is outside of the scope of Theorem 5.1 (since in its current form this result covers only finite model classes), with some extra work and with appropriate modifications one can show that Theorem 5.1 continues to hold for infinite model classes, such as the mentioned logit class.

5.1. Proof Sketch

The purpose of this section is to give a sketch of the proof of Theorem 5.1. For the full proof, see Appendix B. We start by defining the pointwise triangular deviation of f from $q^\star$,

$$\Delta_f^2(s,a) = \frac{(f(s,a) - q^\star(s,a))^2}{f(s,a) + q^\star(s,a)}\,,$$

which is closely related to triangular discrimination (Topsøe, 2000). We can relate Δ_f^2 to the Hellinger distance via the following lemma:

Lemma 5.2. *For all $p, q \in [0,1]$, we have*

$$\frac{1}{4}\frac{(p-q)^2}{p+q} \le \frac{1}{2}\left(\sqrt{p} - \sqrt{q}\right)^2 \le \mathrm{h}^2(p\,\|\,q)\,, \qquad (5)$$

where for $p = q = 0$ we define the left-hand side to be zero and $\mathrm{h}^2(p\,\|\,q) = \frac{1}{2}(\sqrt{p} - \sqrt{q})^2 + \frac{1}{2}(\sqrt{1-p} - \sqrt{1-q})^2$ is the squared Hellinger distance.

The proof of Lemma 5.2 is deferred to Appendix A.1. The idea to relate the pointwise triangular deviation to the squared Hellinger distance was first employed by Foster & Krishnamurthy (2021) in analyzing regret bounds for contextual bandits. Our proof can be summarized by the following three main steps, which correspond to the three terms given in Lemma 5.2.

Step 1: Error decomposition The first step in the proof is to decompose the error (or suboptimality gap), $\bar{v}^{\pi_k} - \bar{v}^\star$, into the product of a small-cost term and the pointwise triangular deviation of f from $q^\star$. The analysis in this step is inspired by the proof of Lemma 1 of Foster & Krishnamurthy (2021). We deviate from their analysis to avoid introducing an extra $|\mathcal{A}|$ factor in the bound. We use the performance difference lemma (Lemma B.4), a *multiplicative* Cauchy-Schwarz (Lemma B.5), i.e. for distribution ν

$$\|x - y\|_{1,\nu} \le \|x + y\|_{1,\nu}^{1/2} \cdot \left\| \frac{(x-y)^2}{(x+y)} \right\|_{2,\nu}^{1/2},$$

and an implicit inequality (i.e. Lemma B.7, step $\star$ in the proof of Proposition B.2) to get a small-cost decomposition of the error:

Proposition 5.3. *Let $f : \mathcal{S} \times \mathcal{A} \to [0,\infty)$ and let $\pi = \pi_f$ be a policy that is greedy with respect to f. Define $D_f = \sup_{h \ge 1} \max(\|\Delta_f\|_{2,\nu_{1,h}}, \|\Delta_f\|_{2,\nu_{2,h}})$. Then, it holds that*

$$\bar{v}^\pi - \bar{v}^\star \le \tilde{C}\left(\frac{D_f}{1-\gamma}\sqrt{\bar{v}^\star} + \frac{D_f^2}{(1-\gamma)^2}\right).$$

where $\tilde{C} > 0$ is an absolute constant.

Here $\nu_{1,h}, \nu_{2,h} \in \mathcal{M}_1(\mathcal{S} \times \mathcal{A})$ are appropriately defined distributions, the details of which can be found in Proposition B.2. In summary, step 1 uses the pointwise triangular deviation of f from $q^\star$ to bound the error by the optimal value function $\bar{v}^\star$.

Step 2: Contraction The second step in our proof starts by bounding the pointwise triangular deviation by the Hellinger distance, i.e.

$$\frac{1}{\sqrt{2}}\|\Delta_f\|_{2,\nu} \le \left\|\sqrt{f} - \sqrt{q^\star}\right\|_{2,\nu}$$

where $\nu \in \mathcal{M}_1(\mathcal{S} \times \mathcal{A})$. In Lemma B.13 we next establish that $\mathcal{T}$ is a γ-pseudo-contraction at $q^\star$ with respect to Hellinger distances: For any $f \ge 0$, $\nu \in \mathcal{M}_1(\mathcal{S} \times \mathcal{A})$,

$$\left\|\sqrt{\mathcal{T}f} - \sqrt{\mathcal{T}q^\star}\right\|_{2,\nu} \le \gamma^{1/2}\left\|\sqrt{f} - \sqrt{q^\star}\right\|_{2,\nu'},$$

where ν' is a distribution over state-action pairs.

Contraction arguments have a long history (Bertsekas, 1995; Littman, 1996; Antos et al., 2007; Chen & Jiang, 2019) in the analysis of dynamic programming algorithms that solve MDPs. The novelty is that we needed to bound the pointwise triangular deviation, which led to new analysis with Hellinger distances.

In Lemma B.15, the combination of a standard change of measure argument (that uses the definition of the concentration coefficient C) and a standard contraction argument, we get

$$\left\|\sqrt{f} - \sqrt{q^\star}\right\|_{2,\nu} \le \frac{\sqrt{C}}{1-\sqrt{\gamma}}\left\|\sqrt{f} - \sqrt{\mathcal{T}f}\right\|_{2,\mu}, \qquad (6)$$

where μ is the exploratory distribution in Assumption 3.1. In batch RL, we often want to learn $q^\star$, or a policy whose value is close to $q^\star$. However, when using function approximation FQI, the regressor corresponding to the training data is $\mathcal{T}f$. Therefore, as demonstrated in Equation (6), if we can show the contraction property then we can bound the error between f and $q^\star$ by the error between f and $\mathcal{T}f$. As we will argue shortly, the error between f and $\mathcal{T}f$ goes to zero as the size of the batch dataset grows.

Step 3: Error control/propagation The third step in our proof starts by bounding

$$\left\| \sqrt{f} - \sqrt{\mathcal{T}f} \right\|_{2,\mu} \leq \sqrt{2} \left\| \mathrm{h}^2(f \parallel \mathcal{T}f) \right\|_{1,\mu}^{1/2} .$$

Then by application of Theorem A.3, we have that if f is the minimizer of $\ell_{\log}$ with respect to the batch dataset D_n, then

$$\sqrt{2} \left\| \mathrm{h}^2(f \parallel \mathcal{T}f) \right\|_{1,\mu}^{1/2} \lesssim \frac{2}{\sqrt{n}} ,$$

where we use $\lesssim$ informally to highlight the most salient elements of the inequality. We then combine all the results to control the pointwise triangular deviation in Proposition 5.3, i.e D_f,

$$\bar{v}^{\pi_k} - \bar{v}^\star \lesssim \frac{\sqrt{C\bar{v}^\star}}{(1-\gamma)^2\sqrt{n}} + \frac{C}{(1-\gamma)^4 n} .$$

This provides a sketch for proving Theorem 5.1. In the full proof we also need to control each iterate of FQI-LOG, i.e. the f_j's. We fill in the missing details in our appendix.

6. Numerical Experiments

The goal of our experiments is to provide insights into the benefits of using FQI-LOG for learning a near-optimal policy in batch reinforcement learning. We run our first set of experiments in reinforcement learning with logit models, as *this setting allows us to best compare FQI-LOG to FQI-SQ without other confounding factors*. For these experiments, we used two standard control tasks; "mountain car" and "inverted pendulum". The tasks are set up as fixed horizon episodic problems where at the end of an episode, if the goal is met no cost is incurred, otherwise a cost of one is incurred. The two environments differ in that in one of them the goal region is small, in the other the goal region is large. In both tasks, some policies can reach the goal, but many fail. As it is known that the goals in these problems can be met, we expect FQI-LOG to do better than FQI-SQ on these environments.

Our second set of experiments aim at verifying whether the recommendation to switch to log loss transfers to deep RL (DRL), i.e., to more complex function classes, regression methods and environments. For these experiments, we started from the work of Agarwal et al. (2020), who tested various DRL methods, including C51 of (Bellemare et al., 2017), a distributional RL algorithm, which was found to be one of the most capable of the methods tested. As noted in the introduction, Wang et al. (2023) showed a small-cost bound for a distributional RL method; hence, our research question is whether a simpler log-loss based method can compete with these (more complex) distributional RL algorithms. To create a real challenge, we picked the two environments (Asterix and Sequest) from Agarwal et al. (2020) where C51 significantly outperformed DQN-SQ, the DRL version of FQI and copied their setting.

6.1. Aiming for a Goal: Mountain Car

We first evaluate FQI-LOG and FQI-SQ on an episodic sparse cost variant of mountain car with episodes lasting for 800 steps. (While we showed our results for the discounted setting, they are expected to hold in episodic problems as well, with small modifications.) Following Moore (1990), this environment consists of a 2-dimensional continuous state space of $[-1.2, 0.6] \times [-0.07, 0.07]$ and 3 discrete actions; the states represent a position and velocity of an underpowered car that can be accelerated left, right, or not accelerated, until the top of a hill is reached when the dynamics is turned on, and the car remains in place regardless of the actions. The cost is 0 at all timesteps except the last, when a cost of 1 is received if the learner has not reached the hilltop. We consider the undiscounted version of the problem (i.e., $\gamma = 1$). An optimal policy for this setting reaches zero cost if it reliably reaches the top of the hill in 800 steps or less, regardless of the exact time. For η_1, the initial state distribution, we use a Dirac that outs the car at the bottom of the hill with zero velocity with probability one.

The feature vectors assigned to states are 16 dimensional and come from a Fourier basis of order 4, following Konidaris et al. (2011) and Chapter 9 of Sutton & Barto (2018). With this, for time step $h \in [800]$, the estimator uses $\theta_h = (\theta_a^h)_{a \in \mathcal{A}} \in \mathbb{R}^{48}$ to produce the estimate $f_h(s,a) = \sigma(\langle \phi(s), \theta_a^h, \rangle)$, where $\sigma(x) = (1 + \exp(-x))^{-1}$ is the sigmoid function. This variant of FQI-LOG with sigmoid functions is closely related to the logistic temporal-difference learning algorithm proposed in Appendix A of the PhD thesis of Silver (2009). We employ the BFGS method, a quasi-Newton method with no learning rate, to find the minimizer of the losses. For strongly-convex functions, BFGS is known to converge to the global minimum superlinearly (Dennis & Moré, 1974). Finally, each batch dataset is constructed from a set of trajectories collected by running the uniform random policy from the initial state $30,000$ times. We use rejection sampling in order to guarantee each dataset has i trajectories that reach the top of the hill with $i \in \{1, 5, 30\}$. We train both FQI-LOG and FQI-SQ on the same batch datasets with the first

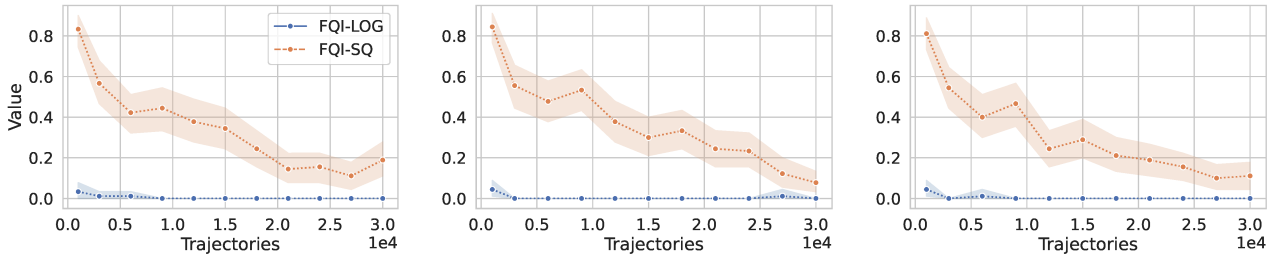

Figure 1. The value of the policy learned by FQI as a function of the size of batch dataset. The results are averaged over 90 independently collected datasets. The figures on the left, middle and right are generated using batch data that contain only 1, 5, and 30 successful trajectories respectively. The standard error of the mean is reported via the shaded region.

$n = [1, 3, 6, 9, 12, 15, 18, 21, 24, 27, 30] \times 10^3$ trajectories in order to study the relationship between the size of the batch dataset and the quality of the policies learned by FQI-LOG and FQI-SQ. We move the successful trajectories to the front of the batch dataset so that they are included for all values of n.

Since BFGS does not require tuning and the same batch datasets are given to FQI-LOG and FQI-SQ, the only variable effecting the performance of the two methods is the loss being minimized. As shown in Figure 1, FQI-LOG accumulates much smaller cost then FQI-SQ using fewer samples, irrespective of the number of trajectories that reach the top of the hill. FQI-LOG is also able to learn a near-optimal policy with only a **single** successful trajectory. In batch RL collecting good trajectories is often expensive, so making efficient use of the few that appear in the batch dataset is an attractive algorithmic feature. As the number of trajectories that reach the top of the hill increases, so too does the performance of FQI-SQ. Since the optimal value on this problem is zero, FQI-LOG is able to learn a near-optimal policy using fewer samples than FQI-SQ.

6.2. Avoiding Failure: Inverted Pendulum

We further evaluate FQI-LOG and FQI-SQ on the inverted pendulum environment (Lagoudakis & Parr, 2003; Riedmiller, 2005), where the goal is to balance an inverted pendulum, by applying forces to it. The state space is two dimensional (angle, angular velocity) and there are three actions (push left, right, no change). The environment dynamics are as described by Lagoudakis & Parr (2003), except that *(i)* when the pendulum falls below horizontal, the state is frozen there and *(ii)* we clip the angular momentum to be in $[-5, 5]$ instead of letting it take arbitrary real values; the clipping is done to facilitate the use of Fourier basis features, which we found works better than the radial basis functions used by Lagoudakis & Parr (2003). As we were using order 4 Fourier features, the parameter space in this case is $4 \times 4 \times 3 = 48$ dimensional. The same logit class was used as in the case of mountain car; for fitting BFGS was used. The cost structure is as follows: The cost of letting the pendulum fall bellow the horizontal is 0. There is no cost otherwise. Again, we know there exist policies that achieves near-zero cost. The discount factor is $\gamma = 0.95$. All datasets are collected by a policy that selects actions uniformly at random until failure (which happens typically after 6 steps) starting from a close to upright, random position. We performed 90 independent trials, each trial consisting of fitting FQI-LOG and FQI-SQ on the same data for $k = 300$ rounds. We then evaluate the learned policies 1000 times on whether the inverted pendulum is still balanced after 3000 steps, starting from a near upright, random position. Results are shown in Figure 2. As with the mountain car experiments, we see that FQI-LOG uses fewer samples than FQI-SQ to learn a good policy.

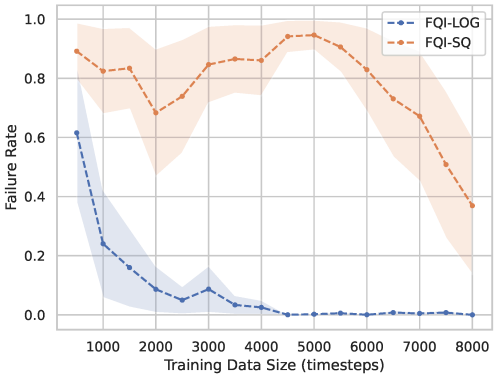

Figure 2. The portion of the time that a policy learned by FQI was able to balance the pendulum for 3000 steps. Results are averaged over 90 independently collected datasets and each learned policy is tested on 1000 initializations. The standard error of the mean is shaded.

6.3. Asterix and Seaquest

As described earlier, we evaluate the deep RL variants of FQI-LOG and FQI-SQ on the Atari 2600 games Asterix and Seaquest (Bellemare et al., 2013), and use the distributional RL algorithm C51 as an additional baseline. We adopt the data and experimental setup of Agarwal et al. (2020). The data consist of five batch datasets for each game, which were collected from independent training runs of a DQN learner (Mnih et al., 2015). Specifically, each batch dataset contains every fourth frame from 200 million frames of training; a frame skip of four and sticky actions (Machado et al., 2018) were used, whereby all actions were repeated four times consecutively and a learner randomly repeated its previous action with probability 0.25.

When the function class $\mathcal{F}$ of FQI-SQ is given by a deep neural network, the algorithm is called DQN. To adapt FQI-LOG to the deep RL setting, we must switch the training loss from ℓ_{sq} to $\ell_{\log}$, and add a sigmoid activation layer to squash the output range to $[0, 1]$. We henceforth refer to these algorithms as DQN-SQ and DQN-LOG, respectively.

The first algorithm to implement a variant of a DQN trained with a form of log-loss is the distributional RL algorithm C51, i.e. categorical DQN (Bellemare et al., 2017). C51 minimizes the *categorical* log-loss across N categories:

$$\ell_{\log,N}(y;\tilde{y}) = \sum_{i=1}^{N} \ell_{\log}(y_i;\tilde{y}_i)\,,$$

for $y, \tilde{y} \in [0,1]^N$. C51 modifies DQN-SQ in the following five ways:

S.1 C51 categorizes the return, i.e. sum of discounted rewards, into 51 "bins", and predicts the probability that the outcome of a state-action pair will fall into each bin, whereas DQN-SQ regresses directly on the returns.

S.2 C51 applies a softmax activation to its output, to normalize the values into a probability distribution over bins, as necessitated by Item **S**.1.

S.3 C51 exchanges ℓ_{sq} for $\ell_{\log,N}$ as the training loss.

S.4 C51 "clips" the targets to the finite interval $[v_{\min}, v_{\max}]$, to enable mapping them into a finite set of bins.

S.5 C51 replaces the Bellman optimality operator $\mathcal{T}$ with a modified "distributional Bellman operator".

For our experiments we clip the targets of DQN-LOG by setting $v_{\min} = 0$ and $v_{\max} = 10$. Since the sigmoid activation of DQN-LOG is a specialization of the softmax activation to the binary case, DQN-LOG implements the changes **S**.3, **S**.2, and **S**.4 to the standard form of DQN-SQ. Clipping the targets introduces a bias which we correct for by similarly clipping the targets of DQN-SQ. Thus our benchmark results for DQN-SQ include the change **S**.4. Clipping the targets of DQN-SQ is novel to this work and improves performance, yielding a stronger baseline. We include a comparison with the traditional unclipped version of DQN-SQ in Appendix C.

In our implementations of DQN-LOG and DQN-SQ, we use the same hyperparameters reported by Agarwal et al. (2020). Figure 3 shows the undiscounted return as a function of the number of training epochs. On Seaquest, DQN-LOG outperforms DQN-SQ and matches the performance of C51. In Asterix, DQN-LOG performs similarly to DQN-SQ and both get lower return than C51. Overall, our results are inconclusive in this setting in regards to whether switching to log-losses suffices to reproduce the success of C51. However, the experiments confirm that switching to log-loss can be beneficial, as compared to using the squared loss and sometimes this switch alone is sufficient to compete with the more complex C51 algorithm.

7. Related Works

First-order bounds in RL Wang et al. (2023) obtain small-cost bounds for finite-horizon batch RL problems under the distributional Bellman completeness assumption, which is more restrictive than our analogue, Assumption 3.4. Wang et al. (2024) refines the bounds of Wang et al. (2023), showing second-order bounds (which depends on the variance) for the same algorithm. They attribute their small-cost bound to the use of the distributional Bellman operator (Bellemare et al., 2023). However, their proof techniques *only* make use of pessimism (Buckman et al., 2021; Jin et al., 2021) and log-loss in achieving their small-cost bound. The use of pessimism is necessary for their proof in order to control the errors accumulated by use of the distribution Bellman operator during value iteration. We improve upon their work by proposing an efficient algorithm for batch RL that enjoys a similar small-cost bound without using the distributional Bellman operator or pessimism under a *weaker* completeness assumption.

Jin et al. (2020a) and Wagenmaker et al. (2022) obtain regret bounds that scale with the value of the optimal policy. However in their setting, the goal is to maximize reward. Therefore, their bounds only improve upon previous bounds (e.g., Azar et al. 2017; Yang & Wang 2019; Jin et al. 2020b) when the optimal policy accumulates very little reward. These bounds are somewhat vacuous as they only imply that regret is low when the value of the initial policy is already close to the value of the optimal policy, both of which are close to zero. Small-cost bounds give the same rates as small-return bounds, however, they are more attractive as the cost of the initial policy can be high while the cost of the optimal

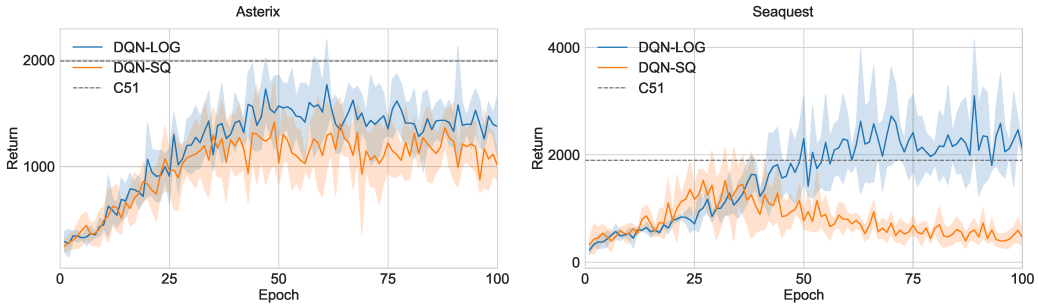

Figure 3. Learning curves on Asterix and Seaquest. The results are averaged over 5 datasets. The shaded regions represent one standard error of the mean. One epoch contains 100k updates.

policy can be low. Small-cost bounds for online learning were given by Lee et al. (2020) for learning in tabular MDPs, while Kakade et al. (2020) derived such results for linear quadratic regulators (LQR). To our knowledge we give the first small-cost bound for batch RL under completeness (Antos et al., 2007; Munos & Szepesvári, 2008).

Small-cost bounds in bandits Several works get small-cost bounds in contextual bandits (Allen-Zhu et al., 2018; Foster & Krishnamurthy, 2021; Olkhovskaya et al., 2023). In their Theorem 3, Foster & Krishnamurthy (2021) show that for batch contextual bandits, playing the greedy policy with respect to a reward function estimated via log-loss enjoys a small-cost bound, whereas in their Theorem 2 they show that if a reward function is estimated via squared loss, the greedy policy fails to achieve a small-cost guarantee. Our main result, Theorem 5.1 can be viewed as the sequel to their Theorem 3. Abeille et al. (2021) show that for stochastic logistic bandits, where the costs/rewards are Bernoulli, the regret can be made to scale with the variance of the optimal arm. This is simultaneously a small-cost and small-return bound. They achieve this bound by employing optimism together with maximum likelihood estimation (MLE). Janz et al. (2024) extend this result and show that this result continues to hold if the reward distribution comes from a "self-concordant", single-parameter family.

Theory on batch RL The theory literature on batch RL has largely focused on proving sample efficiency rates. Chen & Jiang (2019) proved that FQI-SQ gets a rate optimal bound of $1/\sqrt{n}$ when realizability, concentrability and completeness hold. Foster et al. (2021) then show that if one assumes concentrability then completeness is a necessary assumption for sample efficient batch RL. Xie & Jiang (2021) prove that if one uses the stronger notation of concentrability from Munos (2003) then sample efficient batch RL is possible even with only realizability. To our best knowledge, the previous theoretical works on batch RL have only considered algorithms where value estimation used squared loss (e.g., Antos et al., 2007; Farahmand, 2011; Pires & Szepesvári, 2012; Chen & Jiang, 2019; Xie & Jiang, 2021).

Concurrent empirical work The concurrent and independent empirical work of Farebrother et al. (2024) also advocates for log-loss, but they end up with the approach that is used in distributional RL, which reduces regression to multiclass classification. This is not only more complicated than using log-loss (more parameters), but also introduces irreducible bias, whereas our approach avoids this.

8. Conclusions

By proving that FQI-LOG is more sample efficient in MDPs with a small optimal cost $\overline{v}^\star$ than FQI-SQ, we showed that in batch RL the loss function genuinely matters. We believe our result holds generally and can be extended to any batch RL setting where the squared loss has bounded error, such as when pessimistic methods are used. Another intriguing extension would be deriving small-cost bounds in batch RL with only realizability (and a stronger concentrability assumption), perhaps following the analysis of Xie & Jiang (2021). Wagenmaker et al. (2022) get small-return bounds for online RL in linear MDPs. Can we get small-cost bounds in online RL with linear function approximation? When and how? Finally, our mountain car experiments indicate that log-loss might perform well in goal-oriented MDPs (Bertsekas, 1995), where the learner is tasked with reaching some goal and this is possible. However, in this formulation there is no "pressure" for reaching the goal as quickly as possible. It remains to be seen how this could be incorporated in algorithms like ours. (Staying away from failure zones for as long as possible can be easily formulated with the help of discounting.) Our experiments concerning whether switching to the log-loss is sufficient to replicate successes of the more complex C51 distributional RL algorithm were inconclusive. Here, it will be interesting to investigate whether switching to log-loss, but also keeping the losses used by C51 as auxiliary loss can lead to a method that outperforms both C51 and our method.

Acknowledgments

The authors would like to thank Roshan Shariff for pointing out a bug in an earlier version of our proof. Csaba Szepesvári also gratefully acknowledges funding from the Canada CIFAR AI Chairs Program, Amii and NSERC.

Impact Statement

This paper presents work whose goal is to advance the field of reinforcement learning theory. There are many potential societal consequences of our work.

References

Abeille, M., Faury, L., and Calauzènes, C. Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021. [pp. 1 and 9]

Agarwal, R., Schuurmans, D., and Norouzi, M. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2020. [pp. 6, 8, and 24]

Allen-Zhu, Z., Bubeck, S., and Li, Y. Make the minority great again: First-order regret bound for contextual bandits. In *International Conference on Machine Learning (ICML)*, 2018. [p. 9]

Antos, A., Szepesvári, C., and Munos, R. Fitted Q-iteration in continuous action-space MDPs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007. [pp. 1, 3, 4, 5, and 9]

Antos, A., Szepesvári, C., and Munos, R. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 2008. [p. 5]

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2017. [p. 8]

Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. The Arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 2013. [p. 8]

Bellemare, M. G., Dabney, W., and Munos, R. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2017. [pp. 6 and 8]

Bellemare, M. G., Dabney, W., and Rowland, M. *Distributional reinforcement learning*. MIT Press, 2023. [p. 8]

Bertsekas, D. *Dynamic programming and optimal control*. Athena Scientific, 1995. [pp. 5 and 9]

Bertsekas, D. *Reinforcement learning and optimal control*. Athena Scientific, 2019. [p. 2]

Buckman, J., Gelada, C., and Bellemare, M. G. The importance of pessimism in fixed-dataset policy optimization. In *International Conference on Learning Representations (ICLR)*, 2021. [p. 8]

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2019. [pp. 1, 3, 5, and 9]

Dennis, J. E. and Moré, J. J. A characterization of superlinear convergence and its application to quasi-Newton methods. *Mathematics of computation*, 1974. [p. 6]

Ernst, D., Geurts, P., and Wehenkel, L. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research (JMLR)*, 2005. [pp. 1, 3, and 4]

Farahmand, A.-m. *Regularization in reinforcement learning*. PhD thesis, University of Alberta, 2011. [pp. 3, 5, and 9]

Farebrother, J., Orbay, J., Vuong, Q., Taïga, A. A., Chebotar, Y., Xiao, T., Irpan, A., Levine, S., Castro, P. S., Faust, A., Kumar, A., and Agarwal, R. Stop regressing: Training value functions via classification for scalable deep RL. *arXiv preprint arXiv:2403.03950*, 2024. [p. 9]

Foster, D. J. and Krishnamurthy, A. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. [pp. 1, 4, 5, 9, 13, 14, and 16]

Foster, D. J., Krishnamurthy, A., Simchi-Levi, D., and Xu, Y. Offline reinforcement learning: Fundamental barriers for value function approximation. *arXiv preprint arXiv:2111.10919*, 2021. [pp. 3 and 9]

Freund, Y. and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 1997. [p. 5]

Janz, D., Liu, S., Ayoub, A., and Szepesvári, C. Exploration via linearly perturbed loss minimisation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024. [p. 9]

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2020a. [p. 8]

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory (COLT)*, 2020b. [p. 8]

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning (ICML)*, 2021. [p. 8]

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2002. [p. 16]

Kakade, S., Krishnamurthy, A., Lowrey, K., Ohnishi, M., and Sun, W. Information theoretic regret bounds for online nonlinear control. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. [p. 9]

Konidaris, G., Osentoski, S., and Thomas, P. Value function approximation in reinforcement learning using the Fourier basis. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2011. [p. 6]

Lagoudakis, M. G. and Parr, R. Least-squares policy iteration. *The Journal of Machine Learning Research (JMLR)*, 2003. [p. 7]

Lazaric, A., Ghavamzadeh, M., and Munos, R. Finite-sample analysis of least-squares policy iteration. *Journal of Machine Learning Research (JMLR)*, 2012. [p. 5]

Lee, C.-W., Luo, H., Wei, C.-Y., and Zhang, M. Bias no more: High-probability data-dependent regret bounds for adversarial bandits and MDPs. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. [p. 9]

Littman, M. L. *Algorithms for sequential decision-making*. PhD thesis, Brown University, 1996. [p. 5]

Lykouris, T., Sridharan, K., and Tardos, E. Small-loss bounds for online learning with partial information. *Mathematics of Operations Research*, 2022. [p. 5]

Machado, M. C., Bellemare, M. G., Talvitie, E., Veness, J., Hausknecht, M., and Bowling, M. Revisiting the Arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research (JAIR)*, 2018. [p. 8]

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. Human-level control through deep reinforcement learning. *Nature*, 2015. [p. 8]

Moore, A. W. *Efficient memory-based learning for robot control*. PhD thesis, University of Cambridge, 1990. [p. 6]

Munos, R. Error bounds for approximate policy iteration. In *International Conference on Machine Learning (ICML)*, 2003. [pp. 1 and 9]

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research (JMLR)*, 2008. [pp. 5 and 9]

Neu, G. First-order regret bounds for combinatorial semi-bandits. In *Conference on Learning Theory (COLT)*, 2015. [p. 5]

Olkhovskaya, J., Mayo, J., van Erven, T., Neu, G., and Wei, C.-Y. First- and second-order bounds for adversarial linear contextual bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. [p. 9]

Pires, B. Á. and Szepesvári, C. Statistical linear estimation with penalized estimators: An application to reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2012. [pp. 3 and 9]

Riedmiller, M. Neural fitted Q iteration–first experiences with a data efficient neural reinforcement learning method. In *European Conference on Machine Learning (ECML)*, 2005. [pp. 3, 4, and 7]

Silver, D. *Reinforcement learning and simulation-based search in computer Go*. PhD thesis, University of Alberta, 2009. [p. 6]

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018. [p. 6]

Szepesvári, C. *Algorithms for reinforcement learning*. Morgan and Claypool, 2010. [p. 2]

Topsøe, F. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on Information Theory*, 2000. [p. 5]

Wagenmaker, A. J., Chen, Y., Simchowitz, M., Du, S., and Jamieson, K. First-order regret in reinforcement learning with linear function approximation: A robust estimation approach. In *International Conference on Machine Learning (ICML)*, 2022. [pp. 8 and 9]

Wang, K., Zhou, K., Wu, R., Kallus, N., and Sun, W. The benefits of being distributional: Small-loss bounds for reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. [pp. 1, 5, 6, and 8]

Wang, K., Oertell, O., Agarwal, A., Kallus, N., and Sun, W. More benefits of being distributional: Second-order bounds for reinforcement learning. *arXiv preprint arXiv:2402.07198*, 2024. [p. 8]

Xie, T. and Jiang, N. Batch value-function approximation with only realizability. In *International Conference on Machine Learning (ICML)*, 2021. [pp. 1 and 9]

Yang, L. and Wang, M. Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning (ICML)*, 2019. [p. 8]

A. Preliminary results

In this section we introduce and prove some elementary inequalities that connect useful metrics on function spaces, and then state a concentration result of Foster & Krishnamurthy (2021) for the log-loss estimator. The concentration result gives a high probability upper bound on the error of the log-loss estimator, as measured by the integrated binary Hellinger loss (defined below). This result is central to our analysis. The elementary inequalities connect the integrated binary Hellinger loss to both the Hellinger distance and the triangular discrimination, and will reduce the analysis of our algorithm to studying the approximation error of its value function estimates $\{f_j\}_{j=1}^k$.

The analysis of FQI-LOG revolves around controlling the Hellinger distance, which is a distance between nonnegative integrable functions. In particular, for λ-integrable functions $f, g \geq 0$, the Hellinger distance between f and g is defined as

$$\mathrm{H}(f,g) = \frac{1}{\sqrt{2}} \left\| \sqrt{f} - \sqrt{g} \right\|_{2,\lambda} .$$

A.1. Some basic inequalities

Given real numbers $p, q \in [0,1]$, we define the *binary Hellinger loss* of p and q as

$$\mathrm{h}^2(p,q) = \frac{1}{2}\left(\sqrt{p} - \sqrt{q}\right)^2 + \frac{1}{2}\left(\sqrt{1-p} - \sqrt{1-q}\right)^2 , \tag{7}$$

and immediately observe that $0 \leq \mathrm{h}^2(p,q) \leq 1$. Note that the *Hellinger distance* between two distributions P and Q over a common domain is defined as $\frac{1}{\sqrt{2}} \left\| \sqrt{p} - \sqrt{q} \right\|_{2,\lambda}$, where $p = dP/d\lambda$ and $q = dQ/d\lambda$ are the densities of P and Q with respect to a dominating distribution λ. Thus the binary Hellinger loss between p and q, $\mathrm{h}^2(p,q)$, is the squared Hellinger distance between Bernoulli distributions with means p and q.

Lemma A.1. *For all $p, q \in [0,1]$, we have*

$$\frac{1}{4}\frac{(p-q)^2}{p+q} \leq \frac{1}{2}\left(\sqrt{p} - \sqrt{q}\right)^2 \leq \mathrm{h}^2(p,q) , \tag{8}$$

where for $p = q = 0$ we define the left-hand side to be zero.

Proof. If $p = q = 0$ then equality holds trivially, and otherwise $(\sqrt{p} + \sqrt{q})^2 \leq 2(p+q)$ implies

$$\frac{(p-q)^2}{4(p+q)} \leq \frac{(p-q)^2}{2(\sqrt{p}+\sqrt{q})^2} = \frac{1}{2}\left(\sqrt{p} - \sqrt{q}\right)^2 \leq \frac{1}{2}\left(\sqrt{p} - \sqrt{q}\right)^2 + \frac{1}{2}\left(\sqrt{1-p} - \sqrt{1-q}\right)^2 = \mathrm{h}^2(p,q) . \qquad \square$$

The next result holds for an extended definition of h^2 that replaces the inputs $p, q \in [0,1]$ with functions $f, g : \mathcal{X} \to [0,1]$. Given such functions, we define $\mathrm{h}^2(f,g) : \mathcal{X} \to [0,1]$ by

$$(\mathrm{h}^2(f,g))(x) = \mathrm{h}^2(f(x), g(x)) , \qquad x \in \mathcal{X} .$$

With this definition in hand, the following is a straightforward corollary of Lemma A.1.

Corollary A.2. *For any distribution ν over the set $\mathcal{X}$ and any measurable functions $f, g : \mathcal{X} \to [0,1]$,*

$$\left\| \frac{f-g}{\sqrt{f+g}} \right\|_{2,\nu} \leq \sqrt{2} \left\| \sqrt{f} - \sqrt{g} \right\|_{2,\nu} \leq 2 \left\| \mathrm{h}^2(f,g) \right\|_{1,\nu}^{1/2} ,$$

where for $f(x) = g(x) = 0$ we define $\frac{f(x)-g(x)}{\sqrt{f(x)+g(x)}} = 0$.

We call the quantity $\left\| \mathrm{h}^2(f,g) \right\|_{1,\nu}$, which appears on the right hand side above, the *integrated binary Hellinger loss* between f and g. Squaring all quantities and dividing through by 4 yields the equivalent inequalities

$$\frac{1}{4}\left\| \frac{f-g}{\sqrt{f+g}} \right\|_{2,\nu}^2 \leq \frac{1}{2}\left\| \sqrt{f} - \sqrt{g} \right\|_{2,\nu}^2 \leq \left\| \mathrm{h}^2(f,g) \right\|_{1,\nu} .$$

In words, the integrated binary Hellinger loss between f and g is lower bounded by their *squared* Hellinger distance, which itself is lower bounded by one quarter of the triangular discrimination between them. The latter is essentially the squared distance between f and g, but rescaled pointwise by $1/\sqrt{f+g}$. Because $|a-b|/\sqrt{a+b} \le \varepsilon$ implies $|a-b| \le \varepsilon\sqrt{a+b}$, we see that a bound on the rescaled distance between two values tightens the bound between them whenever $\sqrt{a+b} < 1$. Exploiting this property is key to our later analysis.

Proof of Corollary A.2. Apply Lemma A.1 pointwise and then integrate both sides of the inequalities over ν, before taking square roots and multiplying by 2 to simplify the constants. □

A.2. Concentration for the log-loss estimator

Fix a set $\mathcal{X}$, which, for the sake of avoiding measurability issues, is assumed to be finite. Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be independent, identically distributed random elements taking values in $\mathcal{X} \times [0,1]$. Let $f^\star$ be the regression function underlying ν: $f^\star(x) = \mathbb{E}[Y_1 \mid X_1 = x]$. Let $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ be a finite set of $[0,1]$-valued functions with domain $\mathcal{X}$. Recall the log-loss estimator:

$$\hat{f}_{\log} = \operatorname*{arg\,min}_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell_{\log}(f(X_i);\, Y_i),$$

where, for $y, y' \in [0,1]$,

$$\ell_{\log}(y;\, y') = y' \log \frac{1}{y} + (1-y') \log \frac{1}{1-y}\,,$$

where we define $0 \log \infty = \lim_{x \to 0} x \log 1/x = 0$. Foster & Krishnamurthy (2021) show the following concentration result for $\hat{f}_{\log}$, which we will need:

Theorem A.3. *Suppose $f^\star \in \mathcal{F}$. Let $D_n = \{(X_i, Y_i)\}_{i=1}^n$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\| \mathrm{h}^2(\hat{f}_{\log}, f^\star) \|_{1,\nu} \le \frac{2 \log(|\mathcal{F}|/\delta)}{n}\,,$$

where ν denotes the common distribution of $X_1, \dots, X_n$.

Proof. The result follows from the last equation on page 24 of the arXiv version of the paper by Foster & Krishnamurthy (2021) with $A = 1$. □

B. Proof of Theorem 5.1

In this section we give the main steps of the proof of Theorem 5.1. For the benefit of the reader, we first reproduce the text of the theorem.

Theorem B.1. *Given a dataset $D_n = \{(S_i, A_i, C_i, S_i')\}_{i=1}^n$ with $n \in \mathbb{N}$ and a finite function class $\mathcal{F} \subseteq [0,1]^{\mathcal{S}\times\mathcal{A}}$ that satisfy Assumptions 3.1, 3.2 and 3.4, it holds with probability $1-\delta$ that the output policy of FQI-LOG after k iterations, $\pi_k = \pi_{f_k}$, satisfies*

$$\bar{v}^{\pi_k} - \bar{v}^\star \le \tilde{C} \left(\frac{1}{(1-\gamma)^2} \sqrt{\frac{\bar{v}^\star C \log\left(|\mathcal{F}|^2/\delta\right)}{n}} + \frac{C \log\left(|\mathcal{F}|^2/\delta\right)}{(1-\gamma)^4 n} + \frac{\gamma^{\frac{k}{2}}}{1-\gamma} + \frac{\gamma^k}{(1-\gamma)^2} \right).$$

where $\tilde{C} > 0$ is an absolute constant.

The proof is reduced to two propositions and some extra calculations. We start by stating the two propositions first. The proofs of these propositions require more steps and will be developed in their own sections, following the proof of the main result, which ends this section.

The first proposition shows that the error of a policy that is greedy with respect to an action-value function $f : \mathcal{S}\times\mathcal{A} \to [0,\infty)$ can be bounded by the triangular discrimination between the action-value function and $q^\star$, the optimal action-value function

in our MDP. To state this proposition, for f as above, we define $\Delta_f : \mathcal{S} \times \mathcal{A} \to [0, \infty)$, the pointwise triangular deviation of f from $q^\star$:

$$\Delta_f(s,a) = \frac{f(s,a) - q^\star(s,a)}{\sqrt{f(s,a) + q^\star(s,a)}}\,, \qquad (s,a) \in \mathcal{S} \times \mathcal{A}\,.$$

To state the proposition, recall that for a distribution η over the states and a stationary policy π, we let $\eta \times \pi$ denote the joint probability distribution over the state-action pairs resulting from first sampling a state $S \sim \eta$ and then an action $A \sim \pi(S)$. With this, the first proposition is as follows:

Proposition B.2. *Let $f : \mathcal{S} \times \mathcal{A} \to [0, \infty)$ and let $\pi = \pi_f$ be a policy that is greedy with respect to f. Define $D_f = \sup_{h \geq 1} \max(\|\Delta_f\|_{2,\eta_h^\pi \times \pi}, \|\Delta_f\|_{2,\eta_h^\pi \times \pi^\star})$. Then, it holds that*

$$\bar{v}^\pi - \bar{v}^\star \leq \frac{22 D_f}{1-\gamma}\sqrt{2\bar{v}^\star} + \frac{512 D_f^2}{(1-\gamma)^2}\,.$$

Recall that above η_h^π is the distribution induced over the states in step h when π is followed from the start state distribution η_1. As expected, the proof uses the performance difference lemma, followed by arguments that relate the stage-wise expected error that arises from the performance difference lemma to the "size" of Δ_f.

When the above proposition is applied to $f = f_k$, the action-value function obtained in the k^{th} iteration of our algorithm, we see that it remains to bound D_{f_k}. The bound will be based on the second proposition:

Proposition B.3. *For any admissible distribution ν over $\mathcal{S} \times \mathcal{A}$ that may also depend on the data D_n, for any $\delta \in (0,1)$, $k \geq 1$, with probability $1-\delta$,*

$$\|\Delta_{f_k}\|_{2,\nu} \leq \sqrt{\frac{32C\log\left(|\mathcal{F}|^2/\delta\right)}{(1-\gamma)^2 n}} + \sqrt{2}\gamma^{\frac{k}{2}}\,, \tag{9}$$

where f_k denotes the value function computed by FQI, Algorithm 1, in step k based on the data D_n.

The proof of this proposition uses *(i)* showing that $\mathcal{T}$ enjoys some contraction properties with respect to appropriately chosen Hellinger distances; *(ii)* using these contraction properties to show that the Hellinger distance between f_k and $q^\star$ is controlled by the Hellinger distances between f_k and $\mathcal{T} f_k$, and then using the results of the previous section to show that these are controlled by the algorithm.

With these two statements in place, the proof the main theorem is as follows:

Proof of Theorem B.1. Fix $k \geq 1$. For $h \geq 1$, let $\eta_h^k = \eta_h^{\pi_k}$, $D_{f_k} = \sup_{h \geq 1} \max(\|\Delta_{f_k}\|_{2,\eta_h^k \times \pi_k}, \|\Delta_{f_k}\|_{2,\eta_h^k \times \pi^\star})$. Since, by definition, π_k is greedy with respect to f_k, we can use Proposition B.2 to get

$$\bar{v}^{\pi_k} - \bar{v}^\star \leq \frac{22\sqrt{2} D_{f_k}}{1-\gamma}\sqrt{\bar{v}^\star} + \frac{512 D_{f_k}^2}{(1-\gamma)^2}\,. \tag{10}$$

It remains to bound D_{f_k}. An application of Proposition B.3 gives that for any $0 < \delta < 1$, with probability $1-\delta$,

$$S := \sup_{\nu \text{ admissible}} \|D_{f_k}\|_{2,\nu} \leq \sqrt{\frac{32C\log\left(|\mathcal{F}|^2/\delta\right)}{(1-\gamma)^2 n}} + \sqrt{2}\gamma^{\frac{k}{2}}\,. \tag{11}$$

Since $\eta_h^k \times \pi_k$ and $\eta_h^k \times \pi^\star$ are admissible, as can be easily seen with an argument similar to that used in the proof of Lemma B.16, it follows that with probability $1-\delta$,

$$D_{f_k} \leq S \leq \sqrt{\frac{32C\log\left(|\mathcal{F}|^2/\delta\right)}{(1-\gamma)^2 n}} + \sqrt{2}\gamma^{\frac{k}{2}}\,. \tag{12}$$

Squaring both sides and using the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we get that the inequality

$$D_{f_k}^2 \leq \frac{64C\log\left(|\mathcal{F}|^2/\delta\right)}{(1-\gamma)^2 n} + 4\gamma^k$$

also holds, on the same event when Equation (12) holds. Plugging these bounds into Equation (10), we get that with probability at least $1-\delta$,

$$\begin{aligned}\bar{v}^{\pi_k} - \bar{v}^\star &\le \frac{22\sqrt{2}D_{f_k}}{1-\gamma}\sqrt{\bar{v}^\star} + \frac{512 D_{f_k}^2}{(1-\gamma)^2} \\ &\le \frac{176}{(1-\gamma)^2}\sqrt{\frac{\bar{v}^\star C \log\left(|\mathcal{F}|^2/\delta\right)}{n}} + \frac{32768 C \log\left(|\mathcal{F}|^2/\delta\right)}{(1-\gamma)^4 n} + \frac{44\gamma^{\frac{k}{2}}}{1-\gamma} + \frac{2048\gamma^k}{(1-\gamma)^2}\,. \end{aligned}$$

□

B.1. An error bound for greedy policies: Proof of Proposition B.2

The analysis in this section is inspired by the proof of Lemma 1 of Foster & Krishnamurthy (2021). We deviate from their analysis to avoid introducing an extra $|\mathcal{A}|$ factor in the bounds.

Additional Notations For any function $g : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ and policy π, define $g(s,\pi) = \sum_{a\in\mathcal{A}}\pi(a|s)g(s,a)$. For any $\nu \in \mathcal{M}_1(S\times A)$ which is an $|\mathcal{S}\times\mathcal{A}|$ dimensional row vector, define $\nu P \in \mathcal{M}_1(\mathcal{S})$ as the distribution obtained over the states by first sampling a state-action pair from ν and then following P. That is, νP is the distribution of $S' \sim P(\cdot|S,A)$ where $(S,A)\sim\nu$. We can think of νP as the distribution we get when P is composed with ν. For any function $f : \mathcal{S}\times\mathcal{A}\to[0,\infty)$, in addition to Δ_f, we also define $\xi_f : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ as

$$\xi_f(s,a) = f(s,a) + q^\star(s,a)\,, \qquad (s,a)\in\mathcal{S}\times\mathcal{A}\,. \tag{13}$$

Recall that $\mathcal{F}$ contains $[0,1]$-valued functions with domain $\mathcal{S}\times\mathcal{A}$ and as such for any $f\in\mathcal{F}$, Δ_f and ξ_f are well-defined.

We start with the performance difference lemma, which is stated without a proof:

Lemma B.4 (Performance Difference Lemma of Kakade & Langford)**.** *For policies $\pi,\bar{\pi} : \mathcal{S}\to\mathcal{M}_1(\mathcal{A})$, we have*

$$\bar{v}^\pi - \bar{v}^{\bar{\pi}} = \sum_{h=1}^\infty \gamma^{h-1}\left\langle \eta_h^\pi,\, q^{\bar{\pi}}(\cdot,\pi) - v^{\bar{\pi}}\right\rangle\,. \tag{14}$$

Proof. See Lemma 6.1 by Kakade & Langford (2002). □

The next lemma upper bounds the one-norm distance between a nonnegative-valued function $f : \mathcal{S}\times\mathcal{A}\to[0,\infty)$ and $q^\star$ in terms of appropriate norms of Δ_f and ξ_f.

Lemma B.5. *For any function $f : \mathcal{S}\times\mathcal{A}\to[0,\infty)$ and distribution $\nu\in\mathcal{M}_1(S\times A)$, we have*

$$\|f - q^\star\|_{1,\nu} \le \|\xi_f\|_{1,\nu}^{1/2}\cdot\|\Delta_f\|_{2,\nu}\,. \tag{15}$$

Proof. We have

$$\begin{aligned}\|f-q^\star\|_{1,\nu} &= \left\|\sqrt{f+q^\star}\cdot\frac{f-q^\star}{\sqrt{f+q^\star}}\right\|_{1,\nu} && (16)\\ &\le \|f+q^\star\|_{1,\nu}^{1/2}\cdot\left\|\frac{(f-q^\star)^2}{f+q^\star}\right\|_{1,\nu}^{1/2} && \text{(Cauchy-Schwarz)}\\ &= \|\xi_f\|_{1,\nu}^{1/2}\cdot\|\Delta_f\|_{2,\nu}\,. \end{aligned}$$

□

Lemma B.6. *Let $f : S\times A\to[0,\infty)$ and let $\pi=\pi_f$ be a policy that is greedy with respect to f and h be a nonnegative integer. Then it holds that*

$$\left\langle \eta_h^\pi,\, q^\star(\cdot,\pi) - v^\star\right\rangle \le \left(\|\xi_f\|_{1,\eta_h^\pi\times\pi}^{1/2} + \|\xi_f\|_{1,\eta_h^\pi\times\pi^\star}^{1/2}\right)\left(\|\Delta_f\|_{2,\eta_h^\pi\times\pi} + \|\Delta_f\|_{2,\eta_h^\pi\times\pi^\star}\right)\,.$$

Proof. We have

$$\begin{aligned}
\langle \eta_h^\pi,\, q^\star(\cdot,\pi) - v^\star \rangle &= \langle \eta_h^\pi,\, q^\star(\cdot,\pi) - q^\star(\cdot,\pi^\star) \rangle && \text{(Defn of } v^\star\text{)} \\
&\le \langle \eta_h^\pi,\, q^\star(\cdot,\pi) - f(\cdot,\pi) + f(\cdot,\pi^\star) - q^\star(\cdot,\pi^\star) \rangle && (f(\cdot,\pi) \le f(\cdot,\pi^\star) \text{ by defn of } \pi) \\
&\le \|q^\star - f\|_{1,\eta_h^\pi \times \pi} + \|f - q^\star\|_{1,\eta_h^\pi \times \pi^\star} \,. && \text{(triangle inequality)}
\end{aligned}$$

Now,

$$\begin{aligned}
&\|q^\star - f\|_{1,\eta_h^{\tilde\pi} \times \tilde\pi} + \|f - q^\star\|_{1,\eta_h^{\tilde\pi} \times \pi^\star} \\
&\quad\le \|\xi_f\|_{1,\eta_h^{\tilde\pi} \times \tilde\pi}^{1/2} \cdot \|\Delta_f\|_{2,\eta_h^{\tilde\pi} \times \tilde\pi} + \|\xi_f\|_{1,\eta_h^{\tilde\pi} \times \pi^\star}^{1/2} \cdot \|\Delta_f\|_{2,\eta_h^{\tilde\pi} \times \pi^\star} && \text{(Lemma B.5)} \\
&\quad\le \left(\|\xi_f\|_{1,\eta_h^{\tilde\pi} \times \tilde\pi}^{1/2} + \|\xi_f\|_{1,\eta_h^{\tilde\pi} \times \pi^\star}^{1/2} \right) \left(\|\Delta_f\|_{2,\eta_h^{\tilde\pi} \times \tilde\pi} + \|\Delta_f\|_{2,\eta_h^{\tilde\pi} \times \pi^\star} \right) . && \square
\end{aligned}$$

Lemma B.7. *For any function $f : \mathcal{S} \times \mathcal{A} \to [0,\infty)$ and distribution $\nu \in \mathcal{M}_1(\mathcal{S} \times \mathcal{A})$, it holds that*

$$\|f + q^\star\|_{1,\nu} \le 4 \|q^\star\|_{1,\nu} + \|\Delta_f\|_{2,\nu}^2 \,. \tag{17}$$

Proof. Let $f \in \mathcal{F}$ be fixed, we have

$$\begin{aligned}
\|f + q^\star\|_{1,\nu} &= \|f - q^\star + q^\star + q^\star\|_{1,\nu} \\
&\le 2 \|q^\star\|_{1,\nu} + \|f - q^\star\|_{1,\nu} && \text{(triangle inequality)} \\
&= 2 \|q^\star\|_{1,\nu} + \left\| \sqrt{f + q^\star}\, \frac{f - q^\star}{\sqrt{f + q^\star}} \right\|_{1,\nu} \\
&\le 2 \|q^\star\|_{1,\nu} + \frac{1}{2} \left\| f + q^\star + \frac{(f - q^\star)^2}{f + q^\star} \right\|_{1,\nu} && (ab \le \tfrac{a^2+b^2}{2} \text{ for } a, b \text{ nonnegative reals}) \\
&\le 2 \|q^\star\|_{1,\nu} + \frac{1}{2} \|f + q^\star\|_{1,\nu} + \frac{1}{2} \|\Delta_f\|_{2,\nu}^2 \,. && \text{(triangle inequality)}
\end{aligned}$$

Rearranging and multiplying through by two gives the statement. $\square$

Lemma B.8. *Let $f : \mathcal{S} \times \mathcal{A} \to [0,\infty)$ and let $\pi = \pi_f$ be a policy that is greedy with respect to f. Define $D_f = \sup_{h \ge 1} \max(\|\Delta_f\|_{2,\eta_h^\pi \times \pi}, \|\Delta_f\|_{2,\eta_h^\pi \times \pi^\star})$. Then, it holds that*

$$\bar v^\pi - \bar v^\star \le 11 D_f \sum_{h=1}^\infty \gamma^{h-1} \|v^\star\|_{1,\eta_h^k}^{1/2} + \frac{28 D_f^2}{1 - \gamma} \,.$$

Proof. Recall that by the performance difference lemma, Lemma B.4, it holds that

$$\bar v^\pi - \bar v^\star = \sum_{h=1}^\infty \gamma^{h-1} \langle \eta_h^\pi,\, q^\star(\cdot,\pi) - v^\star \rangle \,. \tag{18}$$

For the remainder of this proof we fix $h \ge 1$. For the h^{th} term from the above display, we have

$$\begin{aligned}
\langle \eta_h^\pi,\, q^\star(\cdot,\pi) - v^\star \rangle &\le \left(\|\xi_f\|_{1,\eta_h^\pi \times \pi}^{1/2} + \|\xi_f\|_{1,\eta_h^\pi \times \pi^\star}^{1/2} \right) \left(\|\Delta_f\|_{2,\eta_h^\pi \times \pi} + \|\Delta_f\|_{2,\eta_h^\pi \times \pi^\star} \right) && \text{(Lemma B.6)} \\
&\le \left(\sqrt{4 \|q^\star\|_{1,\eta_h^\pi \times \pi} + \|\Delta_f\|_{2,\eta_h^\pi \times \pi}^2} + \sqrt{4 \|q^\star\|_{1,\eta_h^\pi \times \pi^\star} + \|\Delta_f\|_{2,\eta_h^\pi \times \pi^\star}^2} \right) \left(\|\Delta_f\|_{2,\eta_h^\pi \times \pi} + \|\Delta_f\|_{2,\eta_h^\pi \times \pi^\star} \right) . \\
& && \text{(Lemma B.7)}
\end{aligned}$$

Now recall that by definition

$$\max \left\{ \|\Delta_f\|_{2,\eta_h^\pi \times \pi} , \|\Delta_f\|_{2,\eta_h^\pi \times \pi^\star} \right\} \le D_f \,.$$

Hence,

$$
\begin{aligned}
&\langle \eta_h^\pi,\, q^\star(\cdot,\pi) - v^\star \rangle \\
&\quad \le 2D_f \left(\sqrt{4\,\|q^\star\|_{1,\eta_h^\pi \times \pi} + D_f^2} + \sqrt{4\,\|q^\star\|_{1,\eta_h^\pi \times \pi^\star} + D_f^2} \right) && (19) \\
&\quad \le 2D_f \left(2D_f + \sqrt{4\,\|q^\star\|_{1,\eta_h^\pi \times \pi}} + \sqrt{4\,\|q^\star\|_{1,\eta_h^\pi \times \pi^\star}} \right) && (\sqrt{a+b} \le \sqrt{a} + \sqrt{b}) \\
&\quad = 4D_f^2 + 4D_f\,\|q^\star\|_{1,\eta_h^\pi \times \pi}^{1/2} + 4D_f\,\|q^\star\|_{1,\eta_h^\pi \times \pi^\star}^{1/2} \\
&\quad \le 20D_f^2 + \frac{\|q^\star\|_{1,\eta_h^\pi \times \pi} + \|q^\star\|_{1,\eta_h^\pi \times \pi^\star}}{2}\,. && (ab \le \tfrac{a^2+b^2}{2} \text{ for } a, b \text{ nonnegative reals, twice with } a = 4D_f)
\end{aligned}
$$

Using that $v^\star$, $q^\star$ are nonnegative valued and that $v^\star(\cdot) = q^\star(\cdot, \pi^\star)$, we calculate $\langle \eta_h^\pi, v^\star \rangle = \|v^\star\|_{1,\eta_h^\pi} = \|q^\star\|_{1,\eta_h^\pi \times \pi^\star}$. Thus, by the previous display, after rearranging, we get

$$
\|q^\star\|_{1,\eta_h^\pi \times \pi} = \langle \eta_h^\pi,\, q^\star(\cdot,\pi) \rangle \le 40D_f^2 + 3\,\|q^\star\|_{1,\eta_h^\pi \times \pi^\star}\,,
$$

where the equality used the non-negativity of $q^\star$. Plugging this back into the inequality in Equation (19) gives

$$
\begin{aligned}
\langle \eta_h^\pi,\, q^\star(\cdot,\pi) - v^\star \rangle &\le 2D_f \left(\sqrt{4\,\|q^\star\|_{1,\eta_h^\pi \times \pi^\star} + D_f^2} + \sqrt{4\,\|q^\star\|_{1,\eta_h^\pi \times \pi} + D_f^2} \right) && \text{(restating Equation (19))} \\
&\le 2D_f \left(\sqrt{4\,\|q^\star\|_{1,\eta_h^\pi \times \pi^\star} + D_f^2} + \sqrt{160D_f^2 + 12\,\|q^\star\|_{1,\eta_h^\pi \times \pi^\star} + D_f^2} \right) \\
&\le 2D_f \left(2\,\|q^\star\|_{1,\eta_h^\pi \times \pi^\star}^{1/2} + D_f + \sqrt{161} D_f + \sqrt{12}\,\|q^\star\|_{1,\eta_h^\pi \times \pi^\star}^{1/2} \right) && (\sqrt{a+b} \le \sqrt{a} + \sqrt{b}) \\
&\le 11 D_f\,\|q^\star\|_{1,\eta_h^\pi \times \pi^\star}^{1/2} + 28D_f^2\,.
\end{aligned}
$$

Combining this with Equation (18) gives the desired inequality.

□

Lemma B.9. *For any policy $\pi : \mathcal{S} \to \mathcal{M}_1(\mathcal{A})$ we have*

$$
\sum_{h=1}^{\infty} \gamma^{h-1} \sqrt{\langle \eta_h^\pi,\, v^\pi \rangle} \le \frac{2\sqrt{\bar{v}^\pi}}{1-\gamma}\,. \tag{20}
$$

Proof. Notice that

$$
\begin{aligned}
\bar{v}^\pi &= \langle \eta_1,\, v^\pi \rangle \\
&= \langle \eta_1^\pi,\, c(\cdot,\pi) \rangle + \gamma\,\langle \eta_2^\pi,\, c(\cdot,\pi) \rangle + \gamma^2\,\langle \eta_3^\pi,\, c(\cdot,\pi) \rangle + \cdots + \gamma^{h-1}\,\langle \eta_h^\pi,\, v^\pi \rangle \\
&\ge \gamma^{h-1}\,\langle \eta_h^\pi,\, v^\pi \rangle
\end{aligned}
$$

where the inequality follows from the non-negativity of the costs. Simple rearrangement gives

$$
\langle \eta_h^\pi,\, v^\pi \rangle \le \frac{\bar{v}^\pi}{\gamma^{h-1}}\,.
$$

Using this inequality, we get

$$
\sum_{h=1}^{\infty} \gamma^{h-1} \sqrt{\langle \eta_h^\pi,\, v^\pi \rangle} \le \sum_{h=1}^{\infty} \gamma^{h-1} \sqrt{\frac{\bar{v}^\pi}{\gamma^{h-1}}} = \sum_{h=1}^{\infty} \sqrt{\gamma^{h-1} \bar{v}^\pi} \le \frac{2\sqrt{\bar{v}^\pi}}{1-\gamma}\,,
$$

where for the last inequality we used that $1/(1-\sqrt{\gamma}) \le 2/(1-\gamma)$. □

With this we are ready to prove Proposition B.2:

Proof of Proposition B.2. For $h \geq 1$, let $\eta_h = \eta_h^\pi$. Starting from Lemma B.8, we bound $\bar{v}^\pi - \bar{v}^\star$ as follows:

$$\begin{aligned}
\bar{v}^\pi - \bar{v}^\star &\leq 11 D_f \sum_{h=1}^{\infty} \gamma^{h-1} \left\|v^\star\right\|_{1,\eta_h}^{1/2} + 28 D_f^2 \sum_{h=1}^{\infty} \gamma^{h-1} && \text{(Lemma B.8)} \\
&= 11 D_f \sum_{h=1}^{\infty} \gamma^{h-1} \sqrt{\langle \eta_h,\, v^\star \rangle} + \frac{28 D_f^2}{1-\gamma} \\
&\leq 11 D_f \sum_{h=1}^{\infty} \gamma^{h-1} \sqrt{\langle \eta_h,\, v^\pi \rangle} + \frac{28 D_f^2}{1-\gamma} && \text{(Defn of } v^\star\text{)} \\
&\leq \frac{22 D_f \sqrt{\bar{v}^\pi}}{1-\gamma} + \frac{28 D_f^2}{1-\gamma} && \text{(Lemma B.9, } (\star)\text{)} \\
&\leq \frac{22^2 D_f^2}{2(1-\gamma)^2} + \frac{\bar{v}^\pi}{2} + \frac{28 D_f^2}{1-\gamma}\,. && (ab \leq (a^2+b^2)/2,\, a, b \geq 0)
\end{aligned}$$

Rearranging the last inequality obtained, we get

$$\bar{v}^\pi \leq 2\bar{v}^\star + \frac{22^2 D_f^2}{(1-\gamma)^2} + \frac{56 D_f^2}{1-\gamma}\,.$$

Plugging this bound into $(\star)$, we get

$$\begin{aligned}
\bar{v}^\pi - \bar{v}^\star &\leq \frac{22 D_f}{1-\gamma} \sqrt{2\bar{v}^\star + \frac{22^2 D_f^2}{(1-\gamma)^2} + \frac{56 D_f^2}{1-\gamma}} + \frac{28 D_f^2}{1-\gamma} \\
&\leq \frac{22\sqrt{2} D_f}{1-\gamma} \sqrt{\bar{v}^\star} + \frac{22^2 D_f^2}{(1-\gamma)^2} + \frac{165 D_f^2}{(1-\gamma)^{3/2}} + \frac{28 D_f^2}{1-\gamma} && (\sqrt{a+b} \leq \sqrt{a} + \sqrt{b},\, a, b \geq 0 \text{ twice}) \\
&\leq \frac{22\sqrt{2} D_f}{1-\gamma} \sqrt{\bar{v}^\star} + \frac{512 D_f^2}{(1-\gamma)^2}\,. && \square
\end{aligned}$$

B.2. Bounding the triangular deviation between f_k and $q^\star$: Proof of Proposition B.3

As explained earlier, the analysis in this section uses contraction arguments that have a long history in the analysis of dynamic programming algorithms in the context of MDPs. The novelty is that we need to bound the triangular deviation to $q^\star$. As this has been shown to be upper bounded by the Hellinger distance (Corollary A.2), we switch to Hellinger distances and establishes contraction properties of $\mathcal{T}$ with respect to such distances. This required new proofs. The change of measure arguments used in the "error propagation analysis" are standard.

The following lemma, at a high level, establishes that the map $f \mapsto f^\wedge$ ("min-operator") is a non-expansion over the set of nonnegative functions with domain $\mathcal{S} \times \mathcal{A}$ and $\mathcal{S}$, respectively, when these function spaces are equipped with appropriate norms:

Lemma B.10. *Define the policy $\pi_{f,g}(s) = \arg\min_{a \in A} \min\{f(s,a), g(s,a)\}$ and assume that $f, g : \mathcal{S} \times \mathcal{A} \to [0, \infty)$. Then, for any distribution $\eta \in \mathcal{M}_1(\mathcal{S})$, we have that*

$$\left\| \sqrt{f^\wedge} - \sqrt{g^\wedge} \right\|_{2,\eta} \leq \left\| \sqrt{f} - \sqrt{g} \right\|_{2,\eta \times \pi_{f,f'}}\,. \tag{21}$$

Proof. Notice that for two finite sets of reals, $U = \{u_1, \ldots, u_n\}$, $V = \{v_1, \ldots, v_m\}$, with $u_1 = \min U$, $v_j = \min V$, $u_1 \leq v_j$, $j \in [m]$, we have $|\min U - \min V| = v_j - u_1 \leq v_1 - u_1 \leq |u_1 - v_1|$. By taking the square root of all the elements in both U and V, assuming these are nonnegative, we also get that

$$|\sqrt{\min U} - \sqrt{\min V}| \leq \sqrt{v_1} - \sqrt{u_1} \leq |\sqrt{u_1} - \sqrt{v_1}|\,. \tag{22}$$

Hence,

$$\begin{aligned}
\left\|\sqrt{f^\wedge}-\sqrt{g^\wedge}\right\|_{2,\eta}^2 &= \sum_{s\in S}\eta(s)\left(\sqrt{\min_{a\in A} f(s,a)}-\sqrt{\min_{a\in A} g(s,a)}\right)^2 \\
&= \sum_{s\in S}\eta(s)\left(\sqrt{f\left(s,\pi_f(s)\right)}-\sqrt{g\left(s,\pi_g(s)\right)}\right)^2 \\
&\le \sum_{s\in S}\eta(s)\left(\sqrt{f(s,\pi_{f,g}(s))}-\sqrt{g(s,\pi_{f,g}(s))}\right)^2 && \text{(by Equation (22))} \\
&= \sum_{s\in S}\eta(s)\sum_{a\in A}\mathbb{I}\left\{a=\pi_{f,g}(s)\right\}\left(\sqrt{f(s,a)}-\sqrt{g(s,a)}\right)^2 \\
&= \left\|\sqrt{f}-\sqrt{g}\right\|_{2,\eta\times\pi_{f,g}}^2
\end{aligned}$$

where the inequality used the definition of $\pi_{f,g}(s)$. □

We need two more auxiliary lemmas before we can show the desired contraction result for $\mathcal{T}$. The first is an elementary result that shows that for $x \ge 0$, over the nonnegative reals the map $u \mapsto \sqrt{x+u}$ is a nonexpansion:

Lemma B.11. *For any $x, a, b \ge 0$, we have*

$$\left|\sqrt{x+a}-\sqrt{x+b}\right| \le \left|\sqrt{a}-\sqrt{b}\right|. \tag{23}$$

Proof. For $x \ge 0$, let $f(x) = \left|\sqrt{x+a}-\sqrt{x+b}\right|$. Note that the desired inequality is equivalent to that for any $x \ge 0$, $f(x) \le f(0)$. This, it suffices to show that f is a decreasing function over its domain.

Without loss of generality we may assume that $a > b$ (when $a = b$, the inequality trivially holds, and if $a < b$, just relabel a to b and b to a). Hence, $f(x) = \sqrt{x+a}-\sqrt{x+b}$ for *any* $x \ge 0$ by the monotonicity of the square root function. For $x > 0$, f is differentiable. Here, we get

$$f'(x) = \frac{\partial}{\partial x}\left(\sqrt{x+a}-\sqrt{x+b}\right) = -\frac{\left(\sqrt{x+a}-\sqrt{x+b}\right)^2}{2\sqrt{x+a}\sqrt{x+b}\left(\sqrt{x+a}-\sqrt{x+b}\right)} \le 0. \tag{24}$$

Now, since f is continuous over its domain, by the mean-value theorem, f is decreasing over $[0,\infty)$. □

The next result shows that for any probability distribution λ over some set $\mathcal{X}$, the map $g \mapsto \sqrt{\langle \lambda, g\rangle}$ is a nonexpansion from $H^2(\mathcal{X},\lambda)$ to the reals, where $H^2(\mathcal{X},\lambda)$ is the space of nonnegative valued functions over $\mathcal{X}$ equipped with the Hellinger distance $d(g,h) := \|g^{1/2}-h^{1/2}\|_{2,\lambda}$.

Lemma B.12. *Given a random element X taking values in $\mathcal{X}$ and nonnegative-valued functions $g, g' : \mathcal{X} \to [0,\infty)$ such that $g(X)$ and $g'(X)$ are integrable, we have*

$$\left(\sqrt{\mathbb{E}\,g(X)}-\sqrt{\mathbb{E}\,g'(X)}\right)^2 \le \mathbb{E}\left(\sqrt{g(X)}-\sqrt{g'(X)}\right)^2 < \infty. \tag{25}$$

Proof. The result follows by some calculation:

$$\begin{aligned}
\left(\sqrt{\mathbb{E}\,g(X)}-\sqrt{\mathbb{E}\,g'(X)}\right)^2 &= \mathbb{E}\,g(X) - 2\sqrt{\mathbb{E}\,g(X)}\sqrt{\mathbb{E}\,g'(X)} + \mathbb{E}\,g'(X) \\
&\le \mathbb{E}\,g(X) - 2\mathbb{E}\sqrt{g(X)g'(X)} + \mathbb{E}\,g'(X) && \text{(Cauchy-Schwarz)} \\
&= \mathbb{E}\left[g(X) - 2\sqrt{g(X)g'(X)} + g'(X)\right] \\
&= \mathbb{E}\left(\sqrt{g(X)}-\sqrt{g'(X)}\right)^2,
\end{aligned}$$

where the Cauchy-Schwarz step uses that g and g' are nonnegative. Finally, that $\mathbb{E}\left(\sqrt{g(X)}-\sqrt{g'(X)}\right)^2 < \infty$ follows because, from $(a+b)^2 \le 2(a^2+b^2)$, and hence since $g(X)$ and $g'(X)$ are assumed to be integrable, $\mathbb{E}\left(\sqrt{g(X)}-\sqrt{g'(X)}\right)^2 \le 2\mathbb{E}\,g(X) + 2\mathbb{E}\,g'(X) < \infty$ □

With this we are ready to prove that the Bellman optimality operator is a contraction when used over nonnegative functions equipped with Hellinger distances defined with respect to appropriate measures:

Lemma B.13. *For any distribution $\nu \in \mathcal{M}_1(\mathcal{S}\times\mathcal{A})$, and functions $f, g : \mathcal{S}\times\mathcal{A} \to [0,\infty)$ we have*

$$\left\|\sqrt{\mathcal{T}f}-\sqrt{\mathcal{T}g}\right\|_{2,\nu} \le \gamma^{1/2}\left\|\sqrt{f}-\sqrt{g}\right\|_{2,\nu P\times\pi_{f,g}}.$$

Proof. By the definition of $\mathcal{T}$, we have $\mathcal{T}f = c+\gamma P f^\wedge$. Here, P is viewed as an $SA\times S$ matrix, while c and $f^\wedge$ are viewed as S-dimensional vectors where S and A denote the cardinalities of $\mathcal{S}$ and $\mathcal{A}$ respectively. Also recalling that we use $\sqrt{f}$ to denote the elementwise square root of f for f a vector/function, we have

$$\begin{aligned}
\left\|\sqrt{\mathcal{T}f}-\sqrt{\mathcal{T}g}\right\|_{2,\nu}^2 &= \left\|\sqrt{c+\gamma Pf^\wedge}-\sqrt{c+\gamma Pg^\wedge}\right\|_{2,\nu}^2 \\
&\le \left\|\sqrt{\gamma Pf^\wedge}-\sqrt{\gamma Pg^\wedge}\right\|_{2,\nu}^2 && \text{(Lemma B.11 and the defn. of } \|\cdot\|_{2,\nu}) \\
&= \gamma\left\|\sqrt{Pf^\wedge}-\sqrt{Pg^\wedge}\right\|_{2,\nu}^2 \\
&\le \gamma\left\|\sqrt{f^\wedge}-\sqrt{g^\wedge}\right\|_{2,\nu P}^2 && \text{(Lemma B.12)} \\
&\le \gamma\left\|\sqrt{f}-\sqrt{g}\right\|_{2,\nu P\times\pi_{f,g}}^2, && \text{(Lemma B.10)}
\end{aligned}$$

thus finishing the proof. □

The next result is a simple change-of-measure argument:

Lemma B.14. *Let μ,ν be any distributions over $\mathcal{S}\times\mathcal{A}$ and assume that ν is admissible. Then for $p \ge 1$ we have $\|\cdot\|_{p,\nu} \le C^{1/p}\|\cdot\|_{p,\mu}$.*

Proof. For any function $g : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$, we have

$$\begin{aligned}
\|g\|_{p,\nu} &= \left(\sum_{(s,a)\in S\times A}|g(s,a)|^p\nu(s,a)\right)^{1/p} \\
&\le \left(\sum_{(s,a)\in S\times A}|g(s,a)|^p C\mu(s,a)\right)^{1/p} && \text{(Assumption 3.2)} \\
&= C^{1/p}\left(\sum_{(s,a)\in S\times A}|g(s,a)|^p\mu(s,a)\right)^{1/p} = C^{1/p}\|g\|_{p,\mu}.
\end{aligned}$$

□

Lemma B.15. *Let μ,ν be any distributions over $\mathcal{S}\times\mathcal{A}$ and assume that ν is admissible. Then, for any $f, f' : S\times A\to[0,\infty)$ we have*

$$\left\|\sqrt{f}-\sqrt{q^\star}\right\|_{2,\nu} \le \sqrt{C}\left\|\sqrt{f}-\sqrt{\mathcal{T}f'}\right\|_{2,\mu} + \sqrt{\gamma}\left\|\sqrt{f'}-\sqrt{q^\star}\right\|_{2,\nu P\times\pi_{f',q^\star}}.$$

and $\left\|\sqrt{f}-\sqrt{q^\star}\right\|_{2,\nu} \le \frac{\sqrt{C}}{1-\sqrt{\gamma}}\left\|\sqrt{f}-\sqrt{q^\star}\right\|_{2,\mu}$.

Proof. We have

$$\begin{aligned}
\left\|\sqrt{f}-\sqrt{q^\star}\right\|_{2,\nu} &= \left\|\sqrt{f}-\sqrt{\mathcal{T}f'}+\sqrt{\mathcal{T}f'}-\sqrt{\mathcal{T}q^\star}\right\|_{2,\nu} && (q^\star=\mathcal{T}q^\star)\\
&\le \left\|\sqrt{f}-\sqrt{\mathcal{T}f'}\right\|_{2,\nu}+\left\|\sqrt{\mathcal{T}f'}-\sqrt{\mathcal{T}q^\star}\right\|_{2,\nu} && \text{(triangle inequality)}\\
&\le \sqrt{C}\left\|\sqrt{f}-\sqrt{\mathcal{T}f'}\right\|_{2,\mu}+\sqrt{\gamma}\left\|\sqrt{f'}-\sqrt{q^\star}\right\|_{2,\nu P\times\pi_{f',q^\star}}
\end{aligned}$$

where the last inequality uses Lemmas B.13 and B.14. For the second term let $f'=f$ and $\nu_0=\arg\max_\nu \left\|\sqrt{f}-\sqrt{q^\star}\right\|_{2,\nu}$, then

$$\begin{aligned}
\left\|\sqrt{f}-\sqrt{q^\star}\right\|_{2,\nu_0} &\le \sqrt{C}\left\|\sqrt{f}-\sqrt{q^\star}\right\|_{2,\mu}+\gamma^{1/2}\left\|\sqrt{f}-\sqrt{q^\star}\right\|_{2,\nu_0 P\times\pi_{f,q^\star}}\\
&\le \sqrt{C}\left\|\sqrt{f}-\sqrt{q^\star}\right\|_{2,\mu}+\gamma^{1/2}\left\|\sqrt{f}-\sqrt{q^\star}\right\|_{2,\nu_0}.
\end{aligned}$$

Therefore, $\left\|\sqrt{f}-\sqrt{q^\star}\right\|_{2,\nu}\le\left\|\sqrt{f}-\sqrt{q^\star}\right\|_{2,\nu_0}\le\frac{\sqrt{C}}{1-\sqrt{\gamma}}\left\|\sqrt{f}-\sqrt{q^\star}\right\|_{2,\mu}$. □

Lemma B.16 (Error propagation)**.** *Fix $k\ge 1$ and let $f_0,f_1,\dots,f_k:\mathcal{S}\times\mathcal{A}\to[0,\infty)$ be arbitrary functions such that f_0 takes values in $[0,1]$, ν,μ distributions over $\mathcal{S}\times\mathcal{A}$ and assume that ν is an admissible distribution. Then,*

$$\left\|\sqrt{f_k}-\sqrt{q^\star}\right\|_{2,\nu}\le\gamma^{\frac{k}{2}}+\frac{2\sqrt{C}}{1-\gamma}\max_{1\le\tau\le k}\left\|\sqrt{f_\tau}-\sqrt{\mathcal{T}f_{\tau-1}}\right\|_{2,\mu}.$$

Proof. Define $(\nu_i)_{0\le i\le k}$ via $\nu_k=\nu$ and for $0\le i<k$, let $\nu_i=(\nu_{i+1}P)\times\pi_{f_i,q^\star}$. Note that by assumption, ν_k is admissible. It then follows that ν_i for $0\le i<k$ is also admissible. Indeed, if for some $0\le i<k$, $\pi=(\pi_0,\pi_1,\dots)$ is the nonstationary policy that realizes ν_{i+1} in step $s\ge 0$, $\pi'=(\pi_0,\pi_1,\dots,\pi_s,\pi_{f_i,q^\star},\pi_{s+1},\dots)$ is a policy that realizes ν_i in step $s+1$.

Hence,

$$\begin{aligned}
\left\|\sqrt{f_k}-\sqrt{q^\star}\right\|_{2,\nu} &= \left\|\sqrt{f_k}-\sqrt{q^\star}\right\|_{2,\nu_k} && \text{(definition of } \nu_k)\\
&\le \sqrt{C}\left\|\sqrt{f_k}-\sqrt{\mathcal{T}f_{k-1}}\right\|_{2,\mu}+\sqrt{\gamma}\left\|\sqrt{f_{k-1}}-\sqrt{q^\star}\right\|_{2,\nu_{k-1}}, && \text{(Lemma B.15)}
\end{aligned}$$

where the second inequality uses Lemma B.15 while setting f,f',ν,μ to f_k,f_{k-1},ν_k and μ (the data generating distribution), respectively, and noting that, by definition, $\nu P\times\pi_{f',q^\star}$ of the Lemma is $(\nu_k P)\times\pi_{f_{k-1},q^\star}=\nu_{k-1}$, and that, by assumption, $\nu_k=\nu$ is admissible.

Now, we recurse on the second term of the above display using Lemma B.15:

$$\sqrt{\gamma}\left\|\sqrt{f_{k-1}}-\sqrt{q^\star}\right\|_{2,\nu_{k-1}}\le\sqrt{\gamma C}\left\|\sqrt{f_{k-1}}-\sqrt{\mathcal{T}f_{k-2}}\right\|_{2,\mu}+\gamma\left\|\sqrt{f_{k-2}}-\sqrt{q^\star}\right\|_{2,\nu_{k-2}}$$

where the inequality uses Lemma B.15 while setting f,f',ν,μ to $f_{k-1},f_{k-2},\nu_{k-1}$ and μ (the data generating distribution), respectively, and noting that, by definition, $\nu P\times\pi_{f',q^\star}$ of the Lemma is $(\nu_{k-1}P)\times\pi_{f_{k-2},q^\star}=\nu_{k-2}$, and that, as argued before, ν_{k-1} is admissible.

Continuing this way, and then plugging in back to the first display of the proof, we get

$$\begin{aligned}
\left\|\sqrt{f_k}-\sqrt{q^\star}\right\|_{2,\nu} &\le \sqrt{C}\sum_{j=1}^{k}\gamma^{\frac{k-j}{2}}\left\|\sqrt{f_j}-\sqrt{\mathcal{T}f_{j-1}}\right\|_{2,\mu}+\gamma^{\frac{k}{2}}\left\|\sqrt{f_0}-\sqrt{q^\star}\right\|_{2,\nu_0}\\
&\le \sqrt{C}\underbrace{\sum_{j=1}^{k}\gamma^{\frac{k-j}{2}}\left\|\sqrt{f_j}-\sqrt{\mathcal{T}f_{j-1}}\right\|_{2,\mu}}_{S:=}+\gamma^{\frac{k}{2}},
\end{aligned}$$

where the second inequality holds because by assumption, f_0 takes values in $[0,1]$ and so does $q^\star$. Hence, $(\sqrt{f_0}-\sqrt{q^\star})^2 \le 1$ and thus $\left\|\sqrt{f_0}-\sqrt{q^\star}\right\|_{2,\nu_0} \le 1$.

Now, we bound S defined above:

$$\begin{aligned} S &\le \max_{\tau\in[1,\dots,k]} \left\|\sqrt{f_\tau}-\sqrt{\mathcal{T}f_{\tau-1}}\right\|_{2,\mu} \sum_{j=1}^{k} \gamma^{\frac{k-j}{2}} \\ &\le \frac{2}{1-\gamma} \max_{\tau\in[1,\dots,k]} \left\|\sqrt{f_\tau}-\sqrt{\mathcal{T}f_{\tau-1}}\right\|_{2,\mu} . && (1/(1-\sqrt{\gamma}) \le 2/(1-\gamma)) \end{aligned}$$

Chaining the inequalities finishes the proof. □

Remark B.1. Note that in the proof of the last result it was essential that in the definition of admissibility we allow nonstationary policies.

With this we are ready to prove Proposition B.3:

Proof of Proposition B.3. For the proof let f_t denote the action-value function computed by FQI in step $t = 0, 1, \dots, k$. Recall that by construction $f_0 \in \mathcal{F}$ and that by assumption all functions in $\mathcal{F}$ take values in $[0,1]$. We have

$$\begin{aligned} \|\Delta_{f_k}\|_{2,\nu} &= \left\| \frac{f_k - q^\star}{\sqrt{f_k + q^\star}} \right\|_{2,\nu} && \text{(definition of } \Delta_f) \\ &\le \sqrt{2} \left\| \sqrt{f_k} - \sqrt{q^\star} \right\|_{2,\nu} && \text{(first part of Corollary A.2)} \\ &\le \sqrt{2}\gamma^{\frac{k}{2}} + \frac{2\sqrt{2}\sqrt{C}}{1-\gamma} \max_{1\le\tau\le k} \left\|\sqrt{f_\tau}-\sqrt{\mathcal{T}f_{\tau-1}}\right\|_{2,\mu} && \text{(Lemma B.16, } 0 \le f_0 \le 1) \\ &\le \sqrt{2}\gamma^{\frac{k}{2}} + \frac{4}{1-\gamma} \max_{\tau\in[1,\dots,k]} \left\|\mathrm{h}^2(f_\tau \parallel \mathcal{T}f_{\tau-1})\right\|_{1,\mu}^{1/2} . && \text{(second part of Corollary A.2)} \end{aligned}$$

where in the second inequality we used that by assumption ν is admissible.

For $g : \mathcal{S}\times\mathcal{A} \to [0,1]$, let $\hat{f}_g$ be the function learned by regressing on g via log-loss, i.e.,

$$\hat{f}_g = \arg\min_{f\in\mathcal{F}} \sum_{i=1}^{n} \ell_{\log}\left(f(S_i,A_i);\, C_i + \gamma g^\wedge(S_i')\right) .$$

Note that $f_\tau = \hat{f}_{f_{\tau-1}}$. Hence,

$$\left\|\mathrm{h}^2(f_\tau \parallel \mathcal{T}f_{\tau-1})\right\|_{1,\mu} = \left\|\mathrm{h}^2(\hat{f}_{f_{\tau-1}} \parallel \mathcal{T}f_{\tau-1})\right\|_{1,\mu} \le \max_{g\in\mathcal{F}} \left\|\mathrm{h}^2(\hat{f}_g \parallel \mathcal{T}g)\right\|_{1,\mu} \qquad \text{(because } f_{\tau-1}\in\mathcal{F})$$

Since this applies for any $1 \le \tau \le k$, all that remains is to bound the right-hand side of the last display. We will use Theorem A.3 for this purpose. This result can be applied because, on the one hand, by Assumption 3.1, $\mathbb{E}\left[C_i + \gamma g^\wedge(S_i')|S_i,A_i\right] = \mathcal{T}g\,(S_i,A_i)$, and by Assumption 3.4, $\mathcal{T}g \in \mathcal{F}$ whenever $g\in\mathcal{F}$ and because, again, by Assumption 3.1, (S_i,A_i,C_i,S_{i+1}') are independent, identically distributed random variables for $i = 1,\dots,n$. Thus, Theorem A.3 together with a union bound and recalling that the distribution of (S_i,A_i) is μ gives that for any $0<\delta<1$,

$$\max_{g\in\mathcal{F}} \left\|\mathrm{h}^2(\hat{f}_g, \mathcal{T}g)\right\|_{1,\mu} \le \frac{2\log(|\mathcal{F}|^2/\delta)}{n} .$$

Putting things together, we get that for any fixed $0<\delta<1$, with probability $1-\delta$,

$$\|\Delta_{f_k}\|_{2,\nu} \le \sqrt{\frac{32C\log\left(|\mathcal{F}|^2/\delta\right)}{(1-\gamma)^2 n}} + \sqrt{2}\gamma^{\frac{k}{2}} . \qquad \square$$

C. Experimental details

In this section, we provide additional experimental details and results.

In our Atari experiments, we use the hyperparameters reported in Agarwal et al. (2020), and we do not tune the hyperparameter for our proposed method. The original dataset contains 5 runs of DQN replay data. Each run of DQN replay data contains 50 datasets, and each dataset contains 1 million transitions. Due to the memory constraint, we can not load the entire data. As a result, for each training epoch, we select 5 datasets randomly, subsample a total of 500k transitions from the 5 selected datasets, and perform 100k updates using the 500k transitions.

Figure 4 show the result with clipped losses and unclipped losses. Log-loss consistently outperforms the DQN variants other than C51.

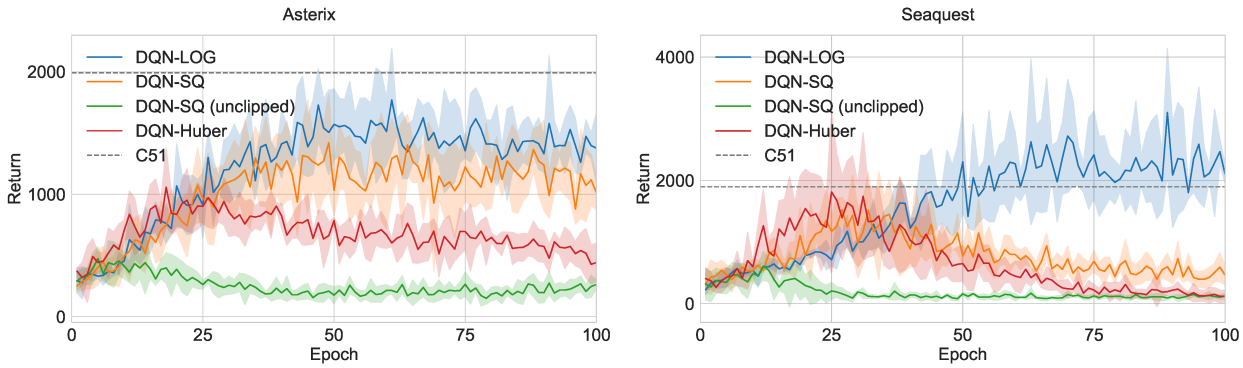

Figure 4. Learning curves on Asterix and Seaquest. The result are averaged over 5 datasets with one standard error. One epoch contains 100k updates.